UNIVERSIDADE NOVA DE LISBOA

Faculdade de Ciências e Tecnologia
Departamento de Matemática

CONSTRUÇÃO DE UMA TARIFA DE RESPONSABILIDADE CIVIL AUTOMÓVEL

Susete Tomás dos Santos

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Matemática e Aplicações – Actuariado, Estatística e Investigação Operacional

Orientador: Prof. Dr. Rui Manuel Rodrigues Cardoso
Co-Orientador: Mestre Maria Teresa Palos Caravina

Lisboa
Dezembro de 2008

AGRADECIMENTOS

Aos meus orientadores, Prof. Doutor Rui Cardoso e Dra. Teresa Caravina, pelo apoio e orientação no desenvolvimento deste trabalho.

À Administração e ao Gabinete Técnico da Companhia de Seguros estudada, pela autorização no uso da informação modelada e pela oportunidade concedida de trabalhar em projectos que contribuíram para o nascimento deste trabalho.

Em especial agradeço à Dra. Tânia Novo e à Dra. Clara Borges, pela amizade sempre demonstrada, pelo imprescindível apoio a todos os níveis e pelos pontos de vista alternativos, sempre enriquecedores.

A todas as pessoas que tiveram a amabilidade de ler este trabalho, pelas suas sugestões.

A todos os meus amigos, pelo apoio emocional e pelo incentivo, essenciais para a conclusão deste trabalho, e por nunca me terem deixado esquecer os meus objectivos.

Finalmente, e em especial, aos meus pais, pelo carinho, apoio, incentivo e confiança sempre presentes em todas as etapas da minha vida.

SUMÁRIO

A actividade seguradora oferece aos seus clientes a transferência para si das eventuais responsabilidades destes, mediante o pagamento de um prémio de seguro, valor que deverá permitir a obtenção de resultados de exploração satisfatórios. Este prémio deverá também ser suficiente para não comprometer a assumpção de responsabilidades em que a seguradora poderá vir a incorrer e, numa visão mais alargada, não deverá comprometer a situação de solvência da empresa.

Estas responsabilidades não são conhecidas no momento em que o prémio é calculado, pelo que as mesmas deverão ser estimadas *à priori*. No entanto, diferentes indivíduos enquadram-se em diferentes classes de risco, pelo que um dos desafios da actividade seguradora é a definição de uma tarifa técnicamente equilibrada e que permita à empresa assegurar o cumprimento das suas responsabilidades, mas que seja também justa e adaptada a cada cliente-tipo.

Nesta dissertação, pretende-se analisar a carteira do Ramo Responsabilidade Civil Automóvel de uma empresa de seguros e, partindo da sua experiência de sinistralidade e da sua estrutura tarifária actual, aferir da necessidade de ajustes à mesma. É ainda proposto analisar a necessidade e viabilidade de construir uma estrutura tarifária alternativa. A via escolhida para a modelização da sinistralidade será a aplicação dos Modelos Lineares Generalizados que, pela sua maleabilidade, têm conhecido aplicações nesta área de trabalho.

PALAVRAS CHAVE: Tarifa, Seguro Responsabilidade Civil Automóvel, Modelos Lineares Generalizados, Modelos de Poisson, Modelos Gamma, Família Exponencial, Estimação pelo Método da Máxima Verosimilhança

ABSTRACT

The Insurance activity assumes the possible responsibilities of their clients, through the payment of an insurance premium, which should be enough to obtain satisfying technical results. This insurance premium should also be enough to assure that the company will be able to respond to its possible future responsibilities and, in a widened vision, it must not compromise the company's solvability.

Such responsibilities are unknown when the premium is calculated, and therefore they must be estimated à priori. However, different individuals fit in different risk classes, and, thus, one of the challenges of the insurance activity is the definition of a technically balanced tariff, which allows the company to assure the fulfilment of it's responsibilities, but which is also a fair and client-adapted tariff.

In this thesis, the objective is to analyse an insurance company's Motor Third Party Liability portfolio and, based on its claims experience and current tariff structure, to evaluate the need to adjust it. It is also considered the need and feasibility of building an alternative tariff structure. Claims history will be modelled through Generalized Linear Models, which, due to its flexibility, have known applications in this line of studies.

KEYWORDS: Tariff, Motor Third Party Liability, Generalized Linear Models, Poisson Models, Gamma Models, Exponential Family, Maximum Likelihood Parameter Estimation

ÍNDICE

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

INTRODUÇÃO

Ao transferir para si as eventuais responsabilidades em que os seus clientes poderão incorrer, e que poderiam não ter possibilidade de suportar a título particular, o sector segurador assume um importante papel social, reforçado pelo facto de o sector ser igualmente um satélite de outras actividades complementares. O sector assume também um importante papel económico, que se traduz num peso relevante do sector na economia nacional.

A actividade seguradora tem sido, no passado recente, confrontada com novos desafios, a nível comercial, económico e legislativo. Atravessando um período macro-económico internacional adverso e num ambiente comercial cada vez mais competitivo, com reflexo nos resultados de 2007 [1] e previsivelmente também nos de 2008, o sector vê-se ainda confrontado com uma maior exigência ao nível da sua gestão de riscos, da assumpção de responsabilidades e dos requisitos de solvabilidade. Citando Pedro Seixas Vale, Presidente do Conselho de Direcção da APS, *"A gestão deste novo ciclo no domínio dos resultados, cobertura de provisões e solvabilidade será um desafio de grande dimensão para todas as seguradoras."* [2]

Neste contexto, a definição de uma tarifa que permita à companhia assegurar o cumprimento das suas responsabilidades futuras e captar o seu público-alvo, evitando também situações de anti-selecção, é um objectivo sempre actual na gestão de uma seguradora. Sendo o prémio estimado à priori, ou seja, com base na sinistralidade conhecida da seguradora, deverão ter-se em consideração os efeitos de diferentes classes de risco nessa sinistralidade.

Nesta dissertação, aborda-se a modelação da tarifa de uma empresa de seguros através dos Modelos Lineares Generalizados, que se aplicam em situações em que pretendemos estimar os efeitos de um determinado factor nas observações, permitindo, assim, alocar o prémio correcto a cada factor de tarifação. Trata-se também de um conjunto de modelos com uma estrutura comum, maleáveis e que permitem o cálculo do erro das estimativas obtidas. Em particular, na componente prática, é analisada uma carteira do Ramo Responsabilidade Civil Automóvel.

Este trabalho está organizado em seis capítulos, no primeiro dos quais, se faz um breve

[1] Que, apesar de positivos, decresceram face a 2006, segundo o Relatório do sector Segurador e Fundos de Pensões, do Instituto de Seguros de Portugal.
[2] www.apseguradores.pt

enquadramento do sector segurador português em termos económico-sociais e ainda do ramo Automóvel, e em particular do ramo Responsabilidade Civil Veículos, tanto no que se refere ao seu peso no sector segurador como aos factores que influenciam a sinistralidade desse ramo. São também abordados os novos desafios que se apresentam ao sector segurador e, mais concretamente, ao ramo Automóvel.

No segundo capítulo, são abordados os conceitos básicos de Prémio, bem como os outros conceitos subjacentes a esta definição, e o conceito de Tarifa.

No terceiro capítulo, abordam-se os conceitos estatísticos que serão necessários ao desenvolvimento do tema em estudo, nomeadamente, a Estimação de Parâmetros, pelo Método da Máxima Verosimilhança, a Família Exponencial de distribuições e o Modelo Linear Clássico. No quarto capítulo, aborda-se a definição dos Modelos Lineares Generalizados e as várias fases da modelação de dados através destes modelos.

No quinto capítulo, aborda-se a modelação da sinistralidade, nas suas componentes Frequência e Severidade, com vista à construção de uma tarifa, utilizando os Modelos Lineares Generalizados, nomeadamente ao nível da selecção da amostra a utilizar e da escolha dos componentes dos Modelos Lineares Generalizados. No último capítulo, é apresentada uma aplicação prática dos modelos definidos neste penúltimo capítulo.

1. O MERCADO SEGURADOR E O SEGURO OBRIGATÓRIO DE RESPONSABILIDADE CIVIL

1.1. MERCADO SEGURADOR PORTUGUÊS – ENQUADRAMENTO ECONÓMICO - SOCIAL

O mercado segurador é um sector com um peso importante na economia portuguesa, como pode ser verificado pelos indicadores constantes da Tabela 1.1.:

	2005	2006	Variação 2006/05	2007	Variação 2007/06
Produto Interno Bruto (PIB) (*)	149.123,50	155.322,60	4,16%	162.756,10	4,79%
População residente (1.000) (**)	10.570	10.599	0,27%	10.618	0,18%
Prémios Emitidos de Seguro Directo do sector segurador (***)	13.223,38	12.729,92	-3,73%	13.748,86	8,00%
Prémios Emitidos / PIB	**8,87%**	**8,20%**	**-7,57%**	**8,45%**	**3,07%**
Prémios/População Residente (1.000)	**1,25**	**1,20**	**-4,00%**	**1,26**	**4,65%**
Resultado Líquido do Exercício do sector segurador (***)	451,70	704,25	55,91%	653,33	-7,23%
Resultado / Prémios	**3,4%**	**5,5%**	**61,96%**	**4,8%**	**-14,11%**
Investimentos Líquidos do sector segurador (***)	39.531,19	44.583,25	12,78%	48.497,26	8,78%
Investimentos Líquidos/ PIB	**26,51%**	**28,70%**	**8,28%**	**29,80%**	**3,81%**

(*) *Fonte: " Relatório do Conselho de Administração" (2005, 2006 e 2007) - Banco de Portugal;* Unidade monetária: Milhões de €
(**) *Fonte:"Estatísticas do Emprego 2007" - Instituto Nacional de Estatística*
(***) Fonte: Instituto de Seguros de Portugal; Unidade monetária: Milhões de €

Tabela 1.1 - O Sector Segurador na economia Portuguesa

Por outro lado, o sector contribui também para a economia e para a sociedade, através da assumpção de custos com sinistros e dos custos com pessoal. Existem ainda sectores directamente dependentes deste sector, nomeadamente os mediadores e corretores de seguros, cuja função é a distribuição de seguros e que o sector segurador remunera, através do pagamento de comissões. Vejamos alguns desses valores:

Unidade monetária: Milhões de €	2005	2006	2007
Custos Brutos com sinistros	6.409,35	7.535,24	9.608,22
Número de empregados	11.836	11.518	11.790
Custo com o pessoal	427,50	429,93	*Não Disponível*
Número de Mediadores e Corretores	38.814	37.466	25.947
Comissões de Mediação e Corretagem	454,44	451,35	*Não Disponível*

Fonte: Instituto de Seguros de Portugal

Tabela 1.2 – Contributo Económico-Social do Sector segurador

Para além da mediação e corretagem, o mercado segurador proporcionou também o desenvolvimento de outras actividades complementares, tais como as empresas resseguradoras, cujo papel é o de seguradoras das companhias de seguros, de forma a reduzir o risco destas; e ainda dos brokers de resseguro, intermediários entre as

resseguradoras e as companhias de seguros. De uma forma indirecta, o sector contribui também para outros sectores, como sejam as oficinas, as empresas de peritagem e de averiguação, as clínicas e as empresas de auditoria, por exemplo.

Os seguros revestem-se também de um carácter social, dado que a companhia de seguros transfere para si a assunção de responsabilidades, que os tomadores de seguro poderiam não ter possibilidade de assumir, caso optassem por auto-seguro. Esta componente social é também assumida pelo legislador, dado alguns seguros serem obrigatórios.

1.2. O RAMO AUTOMÓVEL / O SEGURO DE RESPONSABILIDADE CIVIL

O Ramo Automóvel assume um papel importante no sector segurador. De facto, em 2007, os Prémios Brutos Emitidos do Ramo Automóvel representavam cerca de 44% dos Prémios Brutos Emitidos Não Vida[3]. Neste ramo, ficam garantidos os sinistros ocorridos em consequência de circulação na via pública de veículos terrestres a motor, seus reboques ou semi-reboques. Seguram-se quatro grandes grupos de riscos[4]: Responsabilidade Civil Veículos, que garante, em termos gerais, uma indemnização ou reparação dos danos causados pelo segurado a terceiros; Veículos Terrestres, que garante a reparação dos danos no veículo seguro; Pessoas Transportadas, que garante uma indemnização fixa pelos danos corporais ao condutor e aos passageiros do veículo seguro; e Mercadorias Transportadas, que garante os danos causados aos bens transportados, no caso de transporte colectivo de mercadorias.

Sendo o veículo um dos objectos seguros neste ramo, importa conhecer o parque automóvel português, que se caracteriza por uma grande concentração de veículos ligeiros e de veículos com mais de 5 anos, como pode observar-se de seguida[5]:

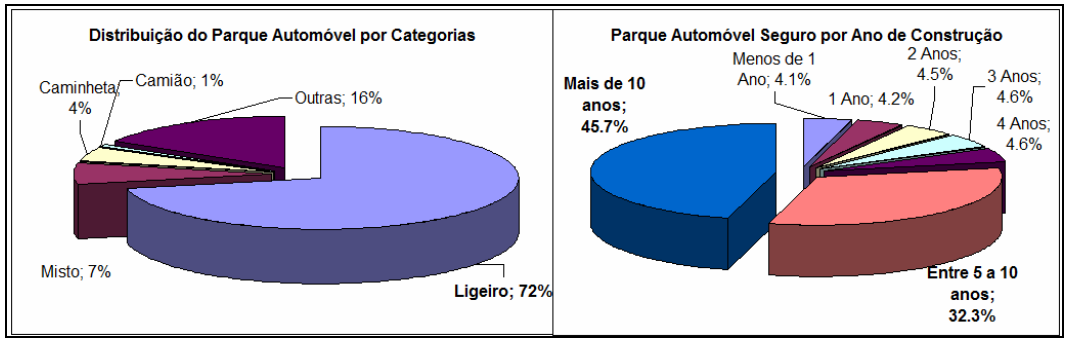


Figura 1-1 - Parque Automóvel Seguro em 2007

[3] Dados Provisórios publicados pelo Instituto de Seguros de Portugal.
[4] Segundo classificação do Instituto de Seguros de Portugal.
[5] Fonte: Instituto de Seguros de Portugal.

A sinistralidade deste ramo tem merecido a atenção não só das companhias de seguros, como também da sociedade civil. Segundo o Relatório de Sinistralidade Rodoviária de 2007, do Observatório de Segurança Rodoviária, o número de acidentes com vítimas diminuiu 1% face a 2006. O número de vítimas mortais manteve-se sensivelmente nos níveis de 2006, mas o número de feridos graves diminui cerca de 11%. A tendência dos últimos anos tem sido de diminuição tanto do número de acidentes, como do número de vítimas, mortais e feridos graves, como pode verificar-se de seguida:

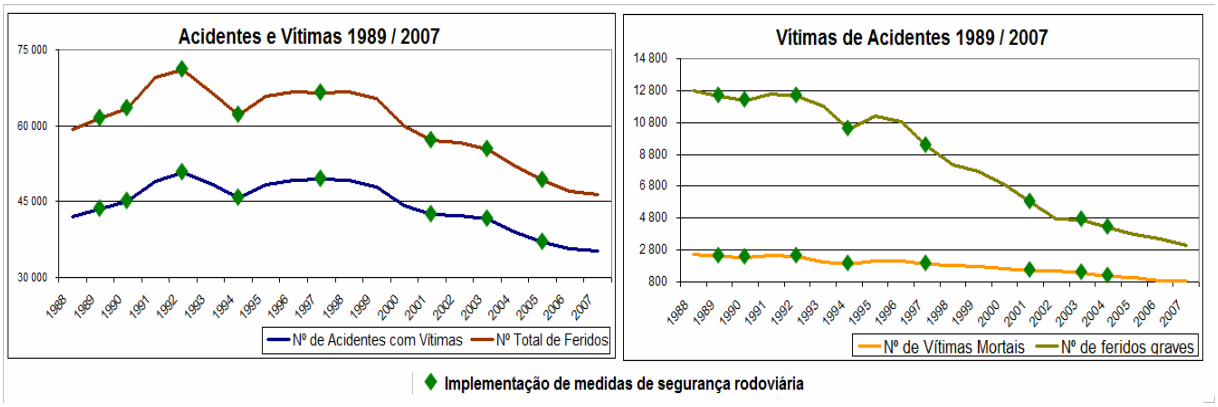


Figura 1-2 - Acidentes Automóvel com vítimas desde 1989 a 2007

Também da observação destes gráficos verificamos que, em geral, a implementação de medidas de segurança rodoviária[6], coincide com uma diminuição da sinistralidade rodoviária.

No entanto, a diminuição do número de acidentes não tem significado uma diminuição dos custos com sinistros assumidos pelas seguradoras, fruto, por um lado, do aumento dos custos de reparação, e por outro lado, de alterações na jurisprudência que têm resultado na condenação das seguradoras ao pagamento de indemnizações mais avultadas. Os custos com sinistros do Ramo Automóvel têm tido a seguinte evolução[7]:

Unidade monetária: Milhões de €	2004	2005	2006	2007
Custos Brutos com Sinistros	1.333,50	1.396,04	1.496,21	Não Disponível

Tabela 1.3 - Evolução dos Custos com Sinistros do Ramo Automóvel

Interessa então analisar os factores que podem estar na origem da sinistralidade automóvel. Para além do incumprimento das regras de segurança e do Código da Estrada, sempre referenciados, o Relatório de Sinistralidade Rodoviária de 2007 apresenta um estudo bastante detalhado sobre este assunto. De seguida, abordar-se-ão os factores mensuráveis e que são, em geral, utilizados pelas seguradoras como factores de tarifação no ramo Automóvel. Note-se no entanto, que estes dados se referem aos acidentes com vítimas, não

[6] Tais como alterações legislativas e campanhas de fiscalização rodoviária.
[7] Fonte: Instituto de Seguros de Portugal.

estando incluídos os acidentes em que apenas se registaram danos materiais[8], pelo que os mesmos devem ser analisados e interpretados, tendo esse aspecto em conta.

Um desses factores é o veículo interveniente no acidente. Verifica-se que, em 2007, os veículos pesados são os que apresentam um maior índice de sinistralidade[9]. Relativamente à idade do veículo, os veículos intervenientes em acidentes têm, maioritariamente, idades compreendidas entre os 0 e os 14 anos (como pode observar-se da figura 1-1, os veículos com esta idade representam 54,3% do parque automóvel português):

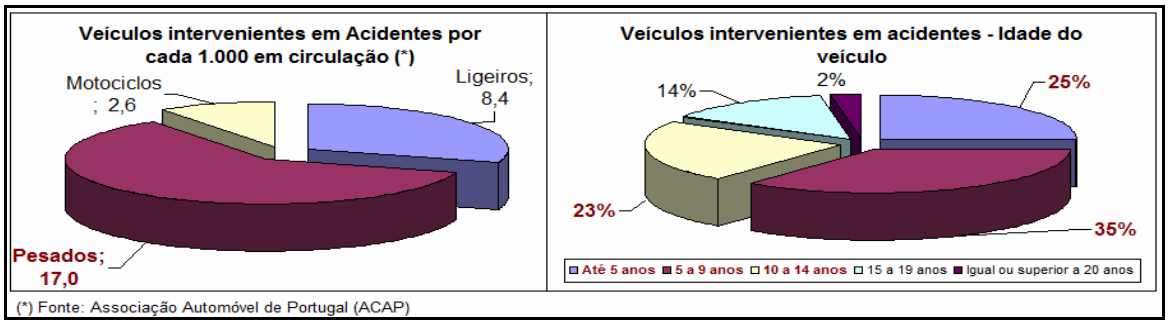


Figura 1-3 - Veículos envolvidos em acidentes com vítimas em 2007

Quanto aos condutores intervenientes em acidentes, verifica-se que, em 2007, a faixa etária com maior sinistralidade é a dos condutores com idade entre os 20 anos e os 34 anos. Verifica-se também que o sexo masculino apresenta um índice de sinistralidade superior ao do sexo feminino. Quanto à antiguidade da carta de condução, a maioria dos condutores tem mais de 11 anos de carta de condução.

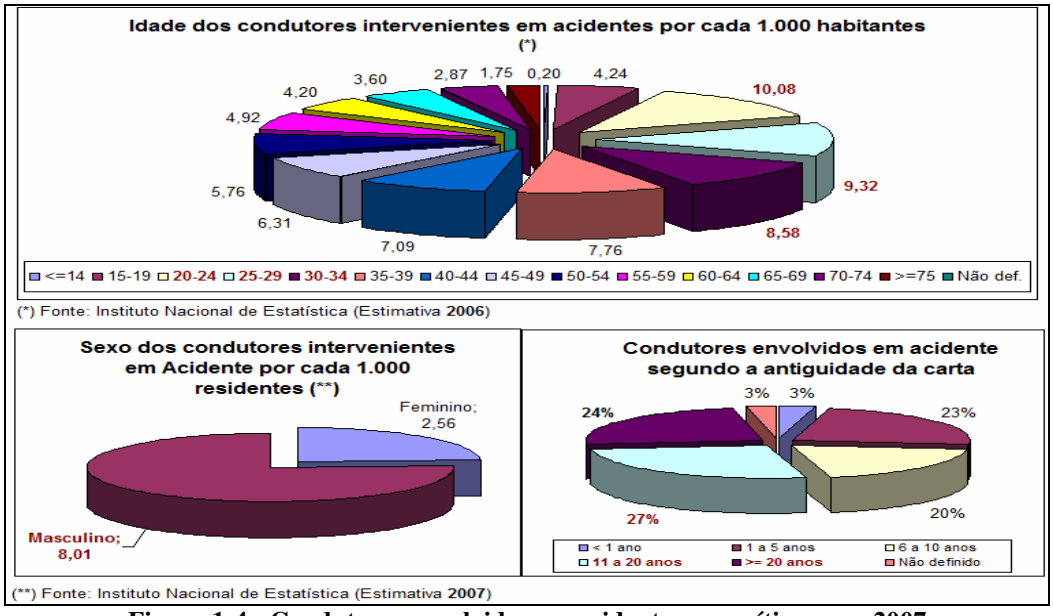


Figura 1-4 - Condutores envolvidos em acidentes com vítimas em 2007

[8] Estando estes últimos incluídos nos Custos com Sinistros suportados pelas seguradoras.
[9] O Índice de Sinistralidade aqui apresentado é o conceito utilizado no Relatório de Sinistralidade Rodoviária, e corresponde ao número de veículos / condutores envolvidos em acidente por cada 1.000 veículos em circulação / residentes.

No que se refere à zona e localização dos acidentes, verifica-se que o maior índice de sinistralidade se verificou na região do Algarve, seguido de Lisboa e da região Centro. Quanto ao tipo de via, a maioria dos acidentes registou-se em arruamentos e estradas nacionais. A maioria dos acidentes regista-se, ainda, dentro das localidades:

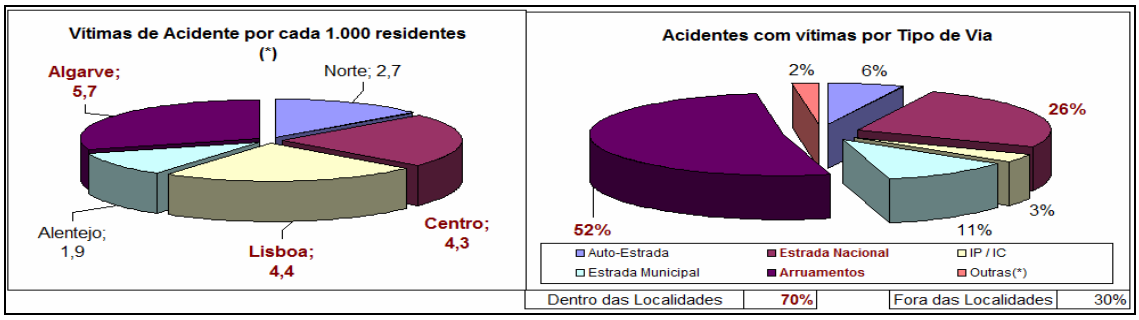


Figura 1-5 - Zona e localização dos acidentes Automóvel com vítimas em 2007

Dentro do Ramo Automóvel, o seguro de Responsabilidade Civil merece particular atenção, dado que se trata de um seguro obrigatório, e cujos Prémios Brutos Emitidos de seguro directo representavam, em 2007, cerca de 66% do Total de Prémios do Ramo Automóvel[10]. A evolução dos prémios e sinistralidade do ramo tem sido a seguinte:

Unidade monetária: Milhões de €	2005	2006	2007
Prémios Brutos Emitidos	1.380,06	1.347,07	1.274,50
Custos Brutos com sinistros	1.090,00	975,00	Não disponível
Rácio S/P[11]	79%	72%	-

Nota: Fonte: Dados até 2006 - Relatório de Mercado, da Associação Portuguesa de Seguradoras / 2007 - Dados Provisórios do Instituto de Seguros de Portugal

Tabela 1.4 - Evolução dos prémios e custos com sinistros de Responsabilidade Civil Veículos

1.3. OS NOVOS DESAFIOS DO SECTOR EM GERAL E DO RAMO AUTOMÓVEL EM PARTICULAR

Embora a análise e prevenção da sinistralidade sejam naturalmente uma das preocupações inerentes ao exercício normal da actividade, o sector segurador em geral e o ramo automóvel, em particular, têm sido também confrontados com novos desafios, a nível comercial, económico e legislativo.

Do ponto de vista comercial, existe uma grande concentração de mercado. Considerando o universo de seguradoras Não Vida e Mistas[12], em 2007, cerca de 67% dos Prémios Brutos Emitidos, concentravam-se em 5 empresas de seguros. Não obstante, o mercado segurador tem conhecido a formação de novos conceitos de distribuição de seguros, como sejam a internet e o telefone, que vieram exigir a adaptação a uma nova realidade comercial, por

[10] Dados Provisórios publicados pelo Instituto de Seguros de Portugal.
[11] Rácio entre o Custo com Sinistros e o montante de Prémios Brutos Emitidos.
[12] Segundo dados do Instituto de Seguros de Portugal, 32 seguradoras Não Vida e Mistas.

parte das seguradoras que utilizam os canais de distribuição tradicionais. De facto, em 2007, os Prémios Brutos Emitidos das duas empresas de seguros que seguem esta filosofia de distribuição de seguros, representavam 3% do Total de Prémios Brutos Emitidos Não Vida e 0,9% dos Prémios Brutos Emitidos das empresas Não Vida e Mistas. Já em 2008, surgiram novas empresas de seguros a operar sob este conceito e viradas para a comercialização de seguros automóvel.

O mercado segurador automóvel é, assim, um mercado bastante competitivo e, numa perspectiva macro-económica, tem sido afectado pela crise económica internacional. O reflexo da conjuntura internacional sente-se, por um lado, ao nível da diminuição do poder de compra dos consumidores e da diminuição da venda de veículos automóveis em Portugal[13], o que naturalmente tem influência directa sobre o mercado segurador. Mais recentemente, a crise nos mercados bolsistas veio também afectar as carteiras de investimentos, situação à qual o sector não foi alheio.

Do ponto de vista da assumpção de responsabilidades, e no que se refere ao ramo automóvel, com a entrada em vigor do Decreto Lei 291/07, de 21 de Agosto, que transpõe a 5ª Directiva Europeia para a legislação portuguesa, estas responsabilidades foram alargadas, por via do aumento do capital mínimo obrigatório e por via de uma maior exigência na celeridade da regularização de sinistros por parte das seguradoras. Também na sequência destas alterações à legislação, foram adoptadas novas regras de cálculo das indemnizações relativas a acidentes que envolvam Danos Corporais.

Um dos recentes desafios colocados à actividade seguradora tem sido o projecto "Solvência II". Embora este projecto se encontre ainda em fase de estudo e de definição e afinação dos critérios de solvência a adoptar, a entrada em vigor dos novos requisitos de capital trará maiores exigências à gestão dos riscos da actividade seguradora.

Neste contexto, o grande desafio do sector é manter o equilíbrio técnico, respondendo a todas as exigências legislativas, ao contexto económico e comercial adverso e ao aumento dos custos com sinistros. Assim, a correcta tarifação de um risco surge como um instrumento importante na resposta a estas condições. Efectivamente, uma Tarifa correcta permite manter um equilíbrio técnico e permite também penetrar em determinados nichos de mercado. No entanto, a elaboração de uma tarifa enfrenta também os seus próprios desafios, pois deve ser suficiente para garantir o pagamento das responsabilidades futuras da empresa de seguros, mas também ser enquadrável no mercado e competitiva.

[13] Segundo dados da Associação Automóvel de Portugal (ACAP), a venda de automóveis em Portugal caiu 3,2% no primeiro semestre de 2008, face ao período homólogo em 2007.

2. TARIFAÇÃO

O contrato de seguro garante a reparação ou pagamento, pela seguradora, dos danos decorrentes de um sinistro que se enquadre nas condições do referido contrato, mediante o pagamento de um prémio, pelo tomador de seguro. Neste contexto, o problema que se coloca é o do cálculo do prémio a cobrar ao tomador, que deve ser suficiente para fazer face ao custo futuro de eventuais sinistros, mas também ser suportável pelo tomador de seguro. Este capítulo começa por apresentar o conceito de prémio, bem como outros conceitos básicos subjacentes ao cálculo do prémio e que serão utilizados no decorrer deste trabalho. Apresentam-se também os Princípios e Modelos de cálculo do prémio mais conhecidos. A segunda secção explora a noção de Tarifa.

2.1. O PRÉMIO, OS PRINCÍPIOS E MODELOS DE CÁLCULO DO PRÉMIO

2.1.1. Conceito e Definições

Um **Sinistro** é "o evento ou série de eventos resultantes de uma mesma causa susceptível de fazer funcionar as garantias do contrato"[14]. O **Prémio de seguro**, é o valor pago pelo tomador de seguro, ao transferir para a seguradora a obrigação de reparar ou pagar os danos decorrentes de um sinistro ocorrido ao abrigo do contrato de seguro.

O Prémio deverá fazer face à sinistralidade futura, sendo assim função do risco, pelo que deverá ter em conta não só o número de sinistros que uma apólice poderá gerar, como também o seu custo, ou seja, tanto a Frequência como a Severidade dos sinistros.

Define-se, **Frequência** como a incidência dos sinistros, expressa em função da exposição ao risco, ou seja: $Frequência = \frac{N^{\circ}\ de\ Sinistros}{N^{\circ}\ de\ Unidades\ \ Expostas\ \ ao\ risco}$

A **Exposição ao risco**, é uma unidade básica do risco, que pretende medir a "quantidade" de risco implícita assumida pela companhia de seguros. Genericamente, utiliza-se o número de apólices como medida. Podemos considerar a Exposição Subscrita, ou seja, as unidades de exposição subscritas durante o período em risco; a Exposição Adquirida – as unidades efectivamente expostas ao risco durante o período em questão, que tem em consideração o período durante o qual as unidades de exposição estiveram em risco; ou a Exposição em

[14] Norma n.º 17/2000-R, de 21 de Dezembro, com as alterações introduzidas pela Norma n.º 13/2005-R, de 18 de Novembro

vigor – as unidades em vigor num determinado momento do tempo, independentemente do período de tempo em que esteve em vigor[15].

A **Severidade de Sinistros** é normalmente expressa pelo Custo Médio com sinistros:

$$Severidade\ de\ Sinitros = \frac{Custo\ com\ Sinistros}{Número\ de\ Sinistros}$$

O **Custo com sinistros** é o montante pago ou pagável aos sinistrados, sendo o **Montante Pago** o montante efectivamente já liquidado. Os montantes que ainda não foram liquidados, mas que se espera vir a pagar, designam-se por **Reserva de sinistros**. Os custos com sinistros podem ainda incluir as **despesas** associadas à regularização do sinistro (excluem-se, no entanto, as despesas administrativas e os custos de investimentos associados).

Para obter o Prémio a cobrar ao segurado, começamos por calcular o **Prémio Puro,** correspondente à perda média por unidade de exposição, obtido através do produto:

$$\Pr émio\ Puro = Frequência * Severidade$$

Adicionando ao Prémio Puro, uma "margem" para despesas administrativas e para desenvolvimento adverso do custo com sinistros, a qual designamos por **Carga de Segurança**, obtemos o **Prémio de Risco**.

Segundo a definição apresentada acima, um Sinistro reveste-se carácter aleatório, pelo que importa, assim, definir formalmente o processo de risco envolvido. O montante total das Indemnizações originadas por um conjunto de riscos, traduz-se por um **processo** composto, $\{S(t)\}_{t\geq 0}$, **das Indemnizações Agregadas**, relativas aos sinistros ocorridos no intervalo de tempo $]0,t]$. Sejam:

- $N(t)$ o Número de Sinistros ocorridos no período de tempo $]0,t]$, sendo $\{N(t)\}_{t\geq 0}$, um processo de contagem;
- X_i, $i = 1, 2, \cdots, N(t)$, o montante da i-ésima indemnização relativa a um sinistro ocorrido em $]0,t]$, sendo $\{X_{(i)}\}_{i=1,2,\cdots,N(t)}$,um conjunto de variáveis aleatórias independentes e identicamente distribuídas e independentes de $N(t)$ e $X_0 \equiv 0$;

Vem:

[15] Esta abordagem exclui as unidades que estiveram a ser cobertas durante um determinado período, mas que entretanto deixaram a carteira.

$$S(t) = \sum_{i=0}^{N(t)} X_i \qquad (2.1)$$

Tendo em conta que $N(t)$ é independente de X_i, o valor esperado das Indemnizações Agregadas originadas pelos sinistros ocorridos num intervalo de tempo $]0,t]$ é:

$$E[S(t)] = E[N(t)] * E[X] \qquad (2.2)$$

Para determinarmos o Prémio Puro, consideramos que $]0,t]$corresponde a uma unidade de tempo. Se $S = S(1)$ representar o montante das indemnizações agregadas ocorridas num determinado período (por exemplo um ano) e $N = N(1)$ representar o número de sinistros ocorridos no mesmo período, vem:

$$\mu_S = E(S) = E(N).E(X) \qquad (2.3)$$

As duas variáveis aleatórias têm naturezas diferentes, pelo que se espera que se comportem de formas diferentes e que, portanto, sigam distribuições de probabilidade diferentes, como será explorado mais à frente.

Tal como já vimos no Capítulo 1, existem vários factores que afectam a sinistralidade[16]. Não sendo possível medir cada risco individualmente, as seguradoras consideram, no cálculo do Prémio Puro, **Factores de Tarifação**, ou seja, factores que representam uma característica do risco a segurar e que pretendem ser uma aproximação do risco o mais possível detalhada e precisa. Estes factores geralmente apresentam ainda uma classificação por **Níveis,** que pretende desagregar o factor de tarifação em riscos mais ou menos gravosos, cobrando um prémio diferenciado em função da gravidade do risco. No entanto, há que ter em atenção que os factores de tarifação e os seus níveis devem ser perfeitamente mensuráveis, de forma a poderem ser aplicados na prática. No caso do seguro de Responsabilidade Civil Automóvel, são exemplos destas aproximações ao risco a idade do condutor, como um dos factores que caracteriza o risco "condutor" ou a categoria do veículo, como um dos factores que caracteriza o risco "veículo".

[16] A análise do capítulo 1 incidiu sobre o ramo automóvel, mas também nos restantes ramos, a sinistralidade depende de mais que um factor.

2.1.2. Princípios de cálculo do Prémio

Um Princípio de cálculo do Prémio é, genericamente, uma regra que permite calcular o Prémio de Risco. Designando-se por P o Prémio de Risco e por θ a Carga de Segurança, sendo θ uma constante positiva, os Princípios de cálculo do Prémio mais conhecidos são:

⇒ **O Princípio do Valor Esperado –** $P = (1+\theta) * \mu_S$,

⇒ **O Princípio do Desvio Padrão –** $P = \mu_S + \theta\sqrt{\sigma_S^2}$,

⇒ **O Princípio da Variância –** $P = \mu_S + \theta\sigma_S^2$.

Os Princípios de cálculo do Prémio possuem várias propriedades desejáveis, sendo que o Princípio do Valor Esperado verifica mais propriedades que os restantes, conforme exposto em *Reis* (2001). Este facto não implica necessariamente que o Princípio do Valor Esperado seja o mais adequado, mas trata-se de um princípio comummente aceite, e em geral este é o princípio mais utilizado.

2.1.3. Modelos de cálculo do Prémio

Seja K uma constante[17] e considere-se que cada função $f_i(x_i)$ representa a influência de cada factor de tarifação, x_i, no custo total com sinistros durante o período em estudo, ou seja, em $S(t)$. Os Modelos de cálculo do Prémio são[18]:

⇒ **Modelo Aditivo** - $P = K * \sum_{i=1}^{k} f_i(x_i)$. Com este modelo, uma alteração no valor de um factor de tarifação implica uma alteração absoluta no valor do prémio.

⇒ **Modelo Multiplicativo -** $P = K * \prod_{i=1}^{k} f_i(x_i)$ Com este modelo, uma alteração no valor de um factor de tarifação implica uma alteração proporcional no valor do prémio, independentemente do valor dos outros factores de tarifação.

⇒ **Modelo Geral -** $P = K * f(x_1, \cdots, x_k)$. Neste modelo, cada grupo de risco é tarifado separadamente.

Na prática, o modelo aditivo e o modelo multiplicativo são os mais utilizados, dada a sua simplicidade de implementação e de interpretação.

[17] Por exemplo, o custo médio com sinistros.
[18] Pitkäen (1975)

Depois de seleccionados os factores de tarifação e de escolhidos o princípio e o modelo de cálculo do prémio, o prémio é apresentado numa Tarifa.

2.2. A TARIFA

A ideia da existência de um "guia" de tarifação, que sirva de orientação ao subscritor, é seguida desde que surgiram os primeiros contratos de seguro, aplicados aos seguros marítimos, em que os prémios de seguro eram baseados nas características de cada navio a ser segurado, nomeadamente na construção e protecção de cada navio, o que resultava numa classificação dada a cada navio e que era anotada e consultada na subscrição de um novo seguro[19].

Hoje em dia, as tarifas têm vários factores em consideração – os factores de tarifação - mas o principal objectivo de uma tarifa continua a ser o de constituir uma medida adequada do risco em que a seguradora incorre ao aceitar subscrever uma determinada apólice. No entanto, a construção de uma tarifa deve ter em conta dois aspectos fundamentais: a adequabilidade do prémio global e a alocação correcta do prémio a cada factor de tarifação.

De facto, o volume global de prémios deve ser adequado, de forma a manter o equilíbrio técnico, ou seja, deve ser suficiente para garantir o pagamento das responsabilidades futuras, das despesas e para fazer face ao desenvolvimento adverso da sinistralidade.

Por outro lado, a aplicação de um prémio igual a todos os riscos não é desejável, pois tal poderia originar uma situação conhecida como "anti-selecção". Ou seja, numa situação em que todos os factores de risco pagam o mesmo prémio, a companhia poderá "atrair" os maus riscos, por outras companhias aplicarem um prémio mais alto a esse factor de risco; podendo, em contrapartida, "desencorajar" a subscrição dos bons riscos, pelas razões contrárias. Não considerar este efeito, pode, assim, levar a que a companhia apenas subscreva bons riscos parcialmente, ou até que não os subscreva, o que, em última análise, pode comprometer a solvência da companhia. A correcta alocação do prémio a cada factor de tarifação pode ainda permitir melhorar a competitividade e rentabilidade em determinados "nichos" de mercado.

O trabalho que se segue debruça-se principalmente sobre este último aspecto, ou seja sobre a correcta alocação de prémio a cada factor de tarifação, através de uma análise estatística dos sinistros conhecidos.

[19] *Mc Clenahan* (2001)

O primeiro objectivo, tanto no que se refere às despesas, como no que se refere ao apuramento da margem para desenvolvimento adverso, a já referida carga de segurança, pode ser atingido através da introdução de factores aplicáveis ao prémio obtido também por ferramentas estatísticas. Esses factores deverão ter em conta a experiência da companhia, bem como factores económicos [20] e outros factores externos [21] . Esta tarifa deverá, posteriormente, ser enquadrada numa perspectiva de mercado, ou seja, de acessibilidade ao cliente e de competitividade com as congéneres.

2.2.1. Modelos de tarifação

Uma das primeiras abordagens sobre a Tarifação foi a construção de tabelas univariadas, em que, para cada factor de tarifação, se apresentam estatísticas simples, tais como a frequência de sinistros e o custo médio de cada factor. Esta análise revela-se útil para uma primeira análise dos dados, permitindo identificar tendências e características específicas da carteira de apólices. No entanto, não é útil para aferir da correcta relação entre os factores de tarifação, nomeadamente ao nível do prémio a alocar a cada um, dado que estas tabelas não consideram interdependências e correlações entre os factores, pelo que os resultados obtidos por esta análise podem conduzir a conclusões incorrectas.

Uma abordagem possível é a seguida pela Teoria da Credibilidade, em que se aplicam factores de credibilidade, ou seja, pesos a um ou mais aspectos da tarifação.

Nos anos 60, os actuários desenvolveram uma técnica que impunha um conjunto de equações sobre os dados observados, as variáveis de tarifação e um conjunto de parâmetros a determinar. O sistema de equações daí resultante é resolvido por um processo iterativo, que procura convergir para a solução óptima. No entanto, apesar de essa solução ser encontrada, esta técnica não fornece uma forma de testar a significância de uma determinada variável no resultado nem uma forma de testar a bondade do ajustamento.

Neste contexto, no início dos anos 70, surgem os Modelos Lineares Generalizados, modelos com uma estrutura generalizada, maleável, que permitem o cálculo do erro das estimativas obtidas e aplicáveis a problemas em que, tal como na construção de uma tarifa, é necessário modelar a relação entre variáveis e estudar a influência que uma ou mais variáveis têm sobre uma outra variável. Estes modelos foram utilizados inicialmente em grupos de investigação restritos, mas o aumento da capacidade computacional e o alargamento da disponibilidade no mercado de software estatístico que permite modelar dados utilizando

20 Por exemplo, a inflação.

21 Por exemplo, alterações na jurisprudência, variações nos custos dos actos médicos, alterações nas políticas de prevenção de riscos (como por exemplo, o aumento de campanhas de prevenção rodoviária).

estes modelos, veio permitir seguir esta nova abordagem na modelação da sinistralidade, nomeadamente no que se refere à área da tarifação.

Os primeiros trabalhos relativos a esta matéria debruçaram-se sobre a modelação da frequência de sinistros e a opção sobre um modelo aditivo ou multiplicativo. Posteriormente, foram também propostos modelos para a modelação da severidade de sinistros.

Vários autores debruçaram-se sobre os modelos a aplicar na modelação da sinistralidade, sendo que os modelos expostos de seguida se baseiam essencialmente nos modelos propostos por *Brockman e Wright* (1992).

3. CONCEITOS ESTATÍSTICOS PRELIMINARES

Para aplicar os Modelos Lineares Generalizados, é necessário estimar parâmetros, pelo que, neste capítulo começa-se, na primeira secção, por apresentar alguns conceitos sobre Estimação de Parâmetros, utilizados neste âmbito, nomeadamente a Estimação pelo Método da Máxima Verosimilhança.

Na segunda secção, apresenta-se um outro conceito sobre o qual se apoiam os Modelos Lineares Generalizados, a Família Exponencial de distribuições, classe de distribuições que reúne várias distribuições conhecidas, numa forma generalizada.

Finalmente, apresenta-se o Modelo Linear Clássico e as razões pelas quais este modelo não era aplicável em muitas situações práticas, o que conduziu ao desenvolvimento dos Modelos Lineares Generalizados.

3.1. ESTIMAÇÃO PONTUAL PELO MÉTODO DE MÁXIMA VEROSIMILHANÇA

O Método da Máxima Verosimilhança tem várias propriedades óptimas, no que se refere à estimação. Sob determinados pressupostos, verdadeiros nos Modelos Lineares Generalizados[22], o estimador obtido é, para grandes amostras:

- Suficiente, ou seja, retira da amostra toda a informação relevante sobre o parâmetro;
- Assintóticamente Consistente, ou seja, quanto maior a amostra, maior a precisão do estimador;
- Eficiente, ou seja, é o estimador que tem a menor variância, de entre todos os estimadores do parâmetro;
- Não Enviesado, ou seja, a Esperança do valor estimado para o parâmetro, $\hat{\alpha}$, é igual ao parâmetro, $\alpha : E[\hat{\alpha}] = \alpha$;
- Tem uma distribuição aproximadamente normal.

Por estas razões, esta técnica é considerada a técnica mais robusta de estimação de parâmetros, ainda que, para amostras pequenas, o estimador obtido possa ser enviesado.

Assim, considere-se uma amostra aleatória (Y_1, Y_2, …, Y_n), em que as observações individuais y_i são independentes, retirada de uma população com uma função densidade de probabilidade f(y, α), a qual depende do vector de parâmetros α (fixo).

[22] *Turkman e Silva* (2000)

A **Verosimilhança** de um conjunto de dados é a probabilidade de obter esse conjunto de dados em particular, dado um modelo de Distribuição de Probabilidade escolhido. Assim, considerando a amostra aleatória (y_1, y_2, …, y_n), a **Função de Verosimilhança** é a função densidade de probabilidade conjunta vista como função do vector de parâmetros desconhecidos, **α,** ou seja, é a função L que satifaz:

$$L(\alpha \mid y) = f(y \mid \alpha) = f(y_1, \alpha) * f(y_2, \alpha) * \ldots * f(y_n, \alpha) \qquad (3.1)$$

Então, L(α,y) representa a verosimilhança do parâmetro α dados os y_1, y_2, …, y_n observados, sendo portanto uma função de α[23].

Determinada a Função de Verosimilhança do modelo em estudo, a questão que se coloca é a obtenção de uma estimativa do valor do parâmetro desconhecido, com vista a descobrir o parâmetro que corresponde à função densidade de probabilidade desejável.

Um método para resolver este problema foi desenvolvido por R. A. Fisher em 1920[24]. Trata-se do **Método de Estimação da Máxima Verosimilhança**, que determina que a função densidade de probabilidade desejável é obtida pelo valor de α que maximiza a Função de Verosimilhança L(α|y). O vector resultante é o que torna o maior possível a probabilidade de se obter a amostra já conhecida, ou seja, A **Estimativa de Máxima Verosimilhança** $\hat{\alpha}$ é a que garante que, sendo $\hat{\hat{\alpha}}$ qualquer outro estimativa de α:

$$L(\hat{\alpha} \mid y) > L(\hat{\hat{\alpha}} \mid y) \qquad (3.2)$$

Em geral, trabalhamos com o logaritmo da função de verosimilhança, dado que o logaritmo é uma função monótona crescente, e, assim, podemos reescrever *(3.3)* da seguinte forma:

$$\ln L(\hat{\alpha} \mid y) > \ln L(\hat{\hat{\alpha}} \mid y) \qquad (3.3)$$

Ou seja, maximizar o logaritmo natural da função de verosimilhança, *ln* (L), produz os mesmos resultados da maximização da função original. Para além disso, a função de verosimilhança é um produto de termos[25], e o logaritmo do produto é a soma do logaritmo dos factores, pelo que em geral é mais simples maximizar *ln* (L).

[23] Sendo assim, é algebricamente semelhante à função densidade de probabilidade, mas difere da mesma, que é função dos Y_i's, dado o parâmetro α, invertendo os papéis do vector de dados Y com o vector de parâmetros α em f(α|y).
[24] *Murteira* (1990)
[25] Por (3.1)

Assumindo que *ln* L é diferenciável, a Estimativa de Máxima Verosimilhança existe, embora possa não ser única. Quando a Estimativa de Máxima Verosimilhança existe, *ln* (L) satisfaz a condição seguinte, designada por **Equação de Verosimilhança:**

$$\boxed{\frac{\partial \ln L(\alpha, y)}{\partial \alpha} = 0} \qquad (3.4)$$

Para além das propriedades desta estimativa já referidas no início deste ponto, outra das propriedades é que se $g(\alpha)$ é uma função dos parâmetros α, então a Estimativa de Máxima Verosimilhança de $g(\alpha)$ é $g(\hat{\alpha})$. Esta propriedade é normalmente designada por **Invariância**, e é útil na prática porque significa que podemos trabalhar com uma função dos parâmetros que seja conveniente para a estimação, e depois utilizar a propriedade de Invariância para obter a Estimativa de Máxima Verosimilhança para os parâmetros.

Na prática, nem sempre é possível obter uma solução analítica para a Estimativa de Máxima Verosimilhança. Nessas situações, a Estimativa de Máxima Verosimilhança tem que ser calculada numericamente, através de algoritmos de optimização linear. No entanto, o algoritmo não garante que se encontre um conjunto de parâmetros que maximize unicamente a função de Log-Verosimilhança, já que a escolha dos parâmetros iniciais do método (recursivo), pode levar a encontrar um máximo local, ou seja, uma solução sub-óptima. No âmbito da existência e unicidade de Estimativas de Máxima Verosimilhança para os Modelos Lineares Generalizados, foram obtidos alguns resultados para casos particulares, que podem ser consultados em Turkman e Silva (2000).

3.2. A FAMÍLIA EXPONENCIAL

3.2.1. Definição Geral

Ao estudarem as propriedades de suficiência estatística de algumas distribuições, *Koopman, Pitman e Darmois*[26] chegam às distribuições da família exponencial, conceito que seria posteriormente introduzido na Estatística por *Fisher*[27].

Diz-se que uma variável aleatória Y **tem distribuição pertencente à família exponencial** se a sua função densidade de probabilidade (ou função de probabilidade, caso Y seja uma variável aleatória discreta) puder escrever-se na forma[28]:

[26] *Murteira* (1990)
[27] *Murteira* (1990)
[28] *Dobson* (2002)

$$f(y;\theta) = s(y)t(\theta)e^{m(y)n(\theta)} \qquad (3.5)$$

sendo $m(.)$, $n(.)$, $s(.)$ e $t(.)$, funções conhecidas.

A expressão acima pode ser reescrita, fazendo $s(y) = \exp(d(y))$ e $t(\theta) = \exp(r(\theta))$, donde se obtém $f(y;\theta) = \exp(d(y)) * \exp(r(\theta)) * \exp(m(y) \cdot n(\theta))$ e:

$$f(y;\theta) = \exp[(m(y).n(\theta) + d(y) + r(\theta)] \qquad (3.6)$$

Se $m(y) = y$, então temos a **distribuição canónica** e $n(\theta)$ é o **parâmetro natural** da distribuição. Podem existir outros parâmetros, integrantes das funções $m(.)$, $n(.)$, $d(.)$ e $r(.)$, que são **parâmetros de dispersão**.

3.2.2. Definição no âmbito dos Modelos Lineares Generalizados

Nos Modelos Lineares Generalizados, como formulado em *McCullagh e Nelder* (1989) temos um conjunto de N variáveis aleatórias independentes, Y_1, …, Y_N , e cada uma com uma distribuição da família exponencial na forma **canónica,** com um **parâmetro de dispersão** conhecido**,** tomando as funções $m(.)$, $n(.)$, $d(.)$ e $r(.)$ a seguinte forma:

$$m(y) = y; \quad n(\theta) = \frac{\theta}{a(\phi)}; \quad r(\theta) = \frac{-b(\theta)}{a(\phi)}; \quad d(y) = c(y,\phi)$$

Ou seja, nesse contexto, pressupõe-se a existência de uma variável resposta pertencente à família exponencial da forma**:**

$$f(y \mid \theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \qquad (3.7)$$

onde que θ e ϕ são parâmetros escalares, e $a(.)$, $b(.)$ e $c(.)$ são funções reais conhecidas. O parâmetro θ, é o **parâmetro canónico**, e está relacionado com a média. O parâmetro ϕ é o **parâmetro de escala** e está relacionado com a variância. Esta relação dos parâmetros com os momentos de primeira e segunda ordem será vista mais à frente. Escolhas diferentes das funções $a(\phi)$, $b(\theta)$ e $c(y,\phi)$ originam diferentes classes de distribuição e uma solução diferente para o parâmetro de escala.

São impostas algumas condições a estas funções:

- $a(\phi)$ é positiva e contínua;

- $b(\theta)$ é diferenciável até à segunda ordem, sendo a segunda derivada uma função positiva;
- $c(y,\phi)$ é independente do parâmetro θ.

De seguida, no âmbito deste trabalho, sempre que se refira a distribuição da família exponencial, a forma utilizada será a (3.7).

Em muitas situações, e, em particular, nos modelos aplicados à tarifação, $a(\phi)$ toma a forma $a(\phi)=\frac{\phi}{\varpi}$, onde ϖ é uma constante conhecida, que atribui um peso ou factor de credibilidade, e que pode estar relacionado com a variabilidade de determinadas observações a modelar ou com a atribuição de menor peso a uma parte dos dados que se sabe ser menos credível.

Neste caso, a função densidade de probabilidade (ou função de probabilidade) definida acima escreve-se na forma:

$$f(y \mid \theta,\phi,\omega)=\exp\left(\frac{\omega}{\phi}(y\theta-b(\theta))+c(y,\phi;\omega)\right) \qquad (3.8)$$

3.2.3. Propriedades das distribuições da família exponencial

As distribuições de cada Y_i são da mesma forma, mas os parâmetros θ_i podem ser diferentes para cada Y_i, ou seja, a distribuição para cada observação Y_i é dada por:

$$f_i(y_i \mid \theta,\phi)=\exp\left(\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+c(y_i,\phi)\right) \qquad (3.9)$$

Assim, cada observação tem um parâmetro canónico θ_i diferente, mas o parâmetro ϕ é igual para todas as observações. Assume-se igualmente que as funções $a(\phi)$, $b(\theta)$ e $c(y,\phi)$ são as mesmas[29] para cada i.

Vejamos agora a relação dos parâmetros θ e ϕ com a média e a variância. Consideremos a Função Geradora de Momentos de f, supondo, sem perda de generalidade, que Y é uma variável aleatória contínua:

[29] Escrevendo na forma (3.8), a função $a(\phi)$ toma a forma $a(\phi)=\phi/\varpi_i$.

$$\varphi_Y(t,\theta,\phi) = E[e^{tY}] = \int_{-\infty}^{+\infty} e^{ty} f(y)\,dy = \int_{-\infty}^{+\infty} \exp(ty).\exp\left[\frac{y\cdot\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]dy =$$

$$= \int_{-\infty}^{+\infty} \exp\left[t\cdot y + \frac{y\cdot\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]dy = \int_{-\infty}^{+\infty} \exp\left[\frac{1}{a(\phi)}\left(a(\phi)\cdot t + \theta\right)\cdot y - \frac{b(\theta)}{a(\phi)} + c(y,\phi)\right]dy =$$

$$= \int_{-\infty}^{+\infty} \exp\left[\frac{1}{a(\phi)}\left(a(\phi)\cdot t + \theta\right)\cdot y + \frac{b\left(a(\phi)\cdot t + \theta\right) - b\left(a(\phi)\cdot t + \theta\right) - b(\theta)}{a(\phi)} + c(y,\phi)\right]dy =$$

$$= \int_{-\infty}^{+\infty} \exp\left[\frac{\left(a(\phi)\cdot t + \theta\right)\cdot y - b\left(a(\phi)\cdot t + \theta\right)}{a(\phi)} + c(y,\phi)\right]dy \cdot \exp\left(\frac{b\left(a(\phi)\cdot t + \theta\right) - b(\theta)}{a(\phi)}\right)$$

E assim, por definição de Função Distribuição:

$$\boxed{\varphi_Y(t,\theta,\phi) = \exp\left(\frac{b\left(a(\phi)\cdot t + \theta\right) - b(\theta)}{a(\phi)}\right)} \qquad (3.10)$$

Por outro lado, temos:

$$\boxed{\psi_Y(t,\theta,\phi) = \ln(\varphi_Y(t,\theta,\phi)) = \frac{b\left(a(\phi)\cdot t + \theta\right) - b(\theta)}{a(\phi)}} \qquad (3.11)$$

Utilizando as propriedades desta função para deduzir os momentos de Y, vem:

$$\psi'_Y(t) = \frac{b'\left(a(\phi)\cdot t + \theta\right)\cdot a(\phi)}{a(\phi)} \Rightarrow \psi'_Y(0) = E[Y] = \mu = b'(\theta) \text{ e}$$

$$\psi''_Y(t) = b''\left(a(\phi)\cdot t + \theta\right)\cdot a(\phi) \Rightarrow \psi''_Y(0) = \mathrm{var}(Y) = b''(\theta)a(\phi)$$

Da mesma forma, para cada Y_i, temos:

$$\boxed{E[Y_i] = \mu_i = b'(\theta_i)} \qquad (3.12)$$

$$\boxed{\mathrm{var}(Y_i) = b''(\theta_i)a(\phi)} \qquad (3.13)$$

A primeira equação define implicitamente θ_i como uma função de μ_i, se for conhecida uma função para $b'(\theta_i)$, pois a primeira equação pode ser resolvida em ordem ao parâmetro canónico θ_i, ou seja, $\theta_i = (b')^{-1}(\mu_i)$.

Por outro lado, a segunda equação indica que a variância de Y_i é uma função do parâmetro canónico, e, portanto, função da média, multiplicada por um termo escalar $a(\phi)$. À primeira

parte da equação, $b''(\theta_i)$, chama-se de **Função Variância**, e, normalmente representa-se por $V(\mu) = b''(\theta_i)$, donde podemos escrever:

$$\boxed{\text{var}(Y_i) = V(\mu) * a(\phi)} \qquad (3.14)$$

Resumindo, dada uma variável aleatória Y_i com distribuição pertencente à família exponencial, esta distribuição tem as seguintes propriedades:

- Cada observação Y_i pertence à mesma classe dentro da família exponencial, mas θ pode variar;
- O parâmetro θ_i é uma função de μ_i;
- Pelas duas propriedades acima, a média de cada observação Y_i pode variar;
- A distribuição fica completamente especificada em termos da sua média e variância, sempre que seja conhecida a função $b(.)$;
- A variância de Y_i é uma função da média (dado que θ_i é função de μ_i).

3.2.4. Distribuições da família exponencial mais conhecidas

Algumas das distribuições mais conhecidas são exemplos de distribuições da família exponencial:

Variáveis de natureza contínua:	Variáveis de natureza discreta:
▪ Distribuição Normal	▪ Distribuição de Poisson
▪ Distribuição Gama	▪ Distribuição Binomial
▪ Distribuição Gaussiana Inversa	

Tabela 3.1 - Família Exponencial: Distribuições mais conhecidas

3.3. O MODELO LINEAR CLÁSSICO

Um **Modelo Linear** procura determinar a relação entre uma variável resposta observada, Y, e determinadas variáveis explicativas ou covariáveis, X. Assumimos que temos os dados na forma $Y_{i=1,...n}$ resultantes da realização de Y em n indivíduos.

Os Modelos Lineares postulam que $Y = \mu + \varepsilon$, assumindo que:

i. O erro tem distribuição Normal, i.e. $\varepsilon \sim N(0, \sigma^2)$

ii. O valor esperado de Y, μ, pode ser escrito como uma combinação linear das variáveis explicativas, X (em forma matricial $X \cdot \beta$, em que β é o vector de parâmetros a estimar, e X a matriz das variáveis explicativas).

Os Modelos Lineares assumem então que:

- As observações individuais são independentes e são Normalmente distribuídas, podendo a média de cada componente variar, mas a sua variância é constante.
- A Média é uma combinação linear das variáveis explicativas da forma $E[Y] = \mu = X \cdot \beta$ e $\eta = X \cdot \beta$, designando-se o preditor linear por η.
- Dado que o vector β é estimado, o mesmo pode ter componentes negativas, para alguma realização das variáveis explicativas. Assim, tanto Y como μ podem tomar qualquer valor real.

Muitos problemas práticos podem formular-se sobre uma amostra aleatória de n observações, e com base na estrutura acima, dada por uma variável resposta e um conjunto de variáveis explicativas. No entanto, o Modelo Linear Clássico tem algumas limitações, que não permitem a sua aplicação a algumas dessas situações práticas, como se segue:

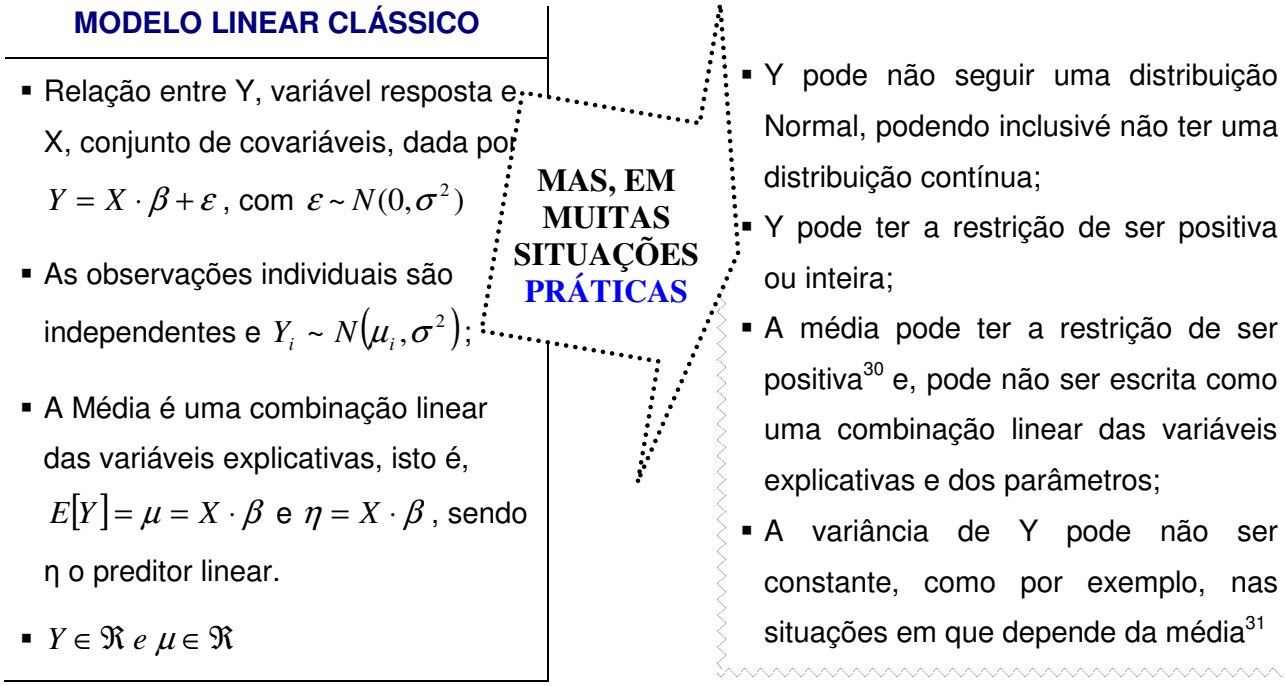


Figura 3-1 Modelo Linear Clássico - Limitações práticas

[30] O que pode não acontecer no Modelo Linear, dado que através da estimação de β podemos obter valores negativos

[31] Um exemplo é a **modelação do número de sinistros**, variável discreta, que toma sempre valores inteiros e em que a média toma sempre valores positivos. Esta variável segue, em muitos casos, uma distribuição de Poisson, em que a variância de cada Y_i é igual à sua média, o que é contrário à assumpção de que a variância é constante.

Para fazer face a estes problemas práticos, vários modelos foram desenvolvidos, aplicáveis a problemas estatísticos que surgem em áreas tão diversas como a medicina, a ecologia, a demografia, a sociologia, entre outras. Alguns desses modelos e as suas aplicações podem ser consultadas em *McCulagah e Nelder* (1989) e *Turkman e Silva* (2000).

4. OS MODELOS LINEARES GENERALIZADOS

Em face da existência de um conjunto de modelos desenvolvidos para fazer face a problemas práticos não explicados pelo Modelo Linear Clássico, *Nelder e Wedderburn* (1972) introduzem uma sintetização desses modelos, os Modelos Lineares Generalizados, que são um conjunto extenso de modelos, dos quais faz parte o Modelo Linear.

Neste capítulo, aborda-se a passagem do Modelo Linear para os Modelos Lineares Generalizados, através da introdução de duas generalizações: a distribuição dos Y_i não tem que ser Normal, podendo ser qualquer distribuição da família exponencial. A média de Y não precisa ser escrita como uma combinação linear do preditor linear, mas essa relação passa a ser expressa através de uma função do preditor linear.

Na segunda secção, é abordada a Modelação dos dados no âmbito dos Modelos Lineares Generalizados, introduzindo-se a Estimação dos parâmetros, a verificação da bondade do ajustamento, através do cálculo dos Resíduos e da *Deviance* e a comparação de modelos.

4.1. DO MODELO LINEAR AOS MODELOS LINEARES GENERALIZADOS

Como referido no capítulo anterior, em muitas situações, o Modelo Linear Clássico não se adequa à realidade prática em estudo e, para fazer face a esses problemas práticos, vários modelos foram desenvolvidos. Em 1972, *Nelder e Wedderburn* apresentam uma teoria que introduz duas generalizações no Modelo Linear, e que agrega todos os modelos até aí desenvolvidos numa estrutura comum (que inclui o próprio Modelo Linear) – os Modelos Lineares Generalizados. Esta teoria unificadora, aliada à expansão da capacidade computacional, trouxe claras vantagens práticas na análise estatística de dados, dado que a existência de uma estrutura comum permite a programação dos pressupostos mais facilmente.

Assim, seguindo as definições utilizadas no capítulo anterior, os **Modelos Lineares Generalizados** procuram estabelecer a relação entre a variável resposta Y_i e o conjunto de variáveis explicativas X_i, caracterizando-se pelo seguinte:

i. Componente Aleatória: As variáveis Y_i são (condicionalmente) independentes e têm uma distribuição pertencente à família exponencial na forma canónica[32], com

[32] Mais concretamente na forna canónica (3.7)

$E[Y_i] = \mu_i = b'(\theta_i)$ para i = 1, …, n e um parâmetro de dispersão ϕ não dependente de i.

ii. Componente estrutural ou sistemática: As variáveis explicativas são combinadas, obtendo-se o preditor linear $\eta = X\beta$

iii. Função de ligação: A relação entre as componentes acima, ou seja entre a média de Y e o preditor linear, é especificada através de uma função de ligação, g, que é diferenciável e monótona, tal que $\eta = g(\mu)$ (ou de forma equivalente $E[Y] = \mu = g^{-1}(\eta)$). Assim, não é necessariamente a média a ser modelada de uma forma linear, mas sim uma transformação adequada da média que é modelada de forma linear.

Ou seja, face ao Modelo Linear são introduzidas duas generalizações, que respondem às condicionantes práticas sentidas na aplicação desse Modelo:

- A distribuição dos Y_i não tem que ser Normal, podendo ser qualquer distribuição da família exponencial, onde se incluem a própria distribuição Normal e também distribuições discretas e com domínio $\Re^+$ e cuja variância não é necessariamente constante;
- A Média de Y não precisa ser escrita como uma combinação linear, mas essa relação passa a ser expressa através de uma função do preditor linear, que mantenha a linearidade e as restrições existentes para a média de Y (como por exemplo, ser positiva).

Para melhor visualizar a "transformação" do Modelo Linear Clássico nos Modelos Lineares Generalizados, podemos reescrever a formulação do Modelo Linear:

GENERALIZAÇÃO

	MODELO LINEAR CLÁSSICO	MODELOS LINEARES GENERALIZADOS
Componente Aleatória	$Y_i \sim N(\mu_i, \sigma^2)$	$Y_i \sim F_{\exp}$, em que $F_{\exp}$ é uma distribuição da família exponencial, com[33] $E[Y_i] = \mu_i = b'(\theta)$ e $\text{var}[Y_i] = b''(\theta).a(\phi) = V(\mu_i).a(\phi)$
Componente estrutural ou sistemática	$\eta = \sum_i^p X_j \beta_j$	$\eta = \sum_i^p X_j \beta_j$
Função de ligação – g(.)	$\eta = \mu$, ou seja, g(.) é a Identidade	$\eta = g(\mu)$, em que g(.) é uma função diferenciável e monótona
Formulação do Modelo	$E[Y] = \mu = \sum_i^p X_j \beta_j$	$E[Y] = g^{-1}(\eta) = g^{-1}\left(\sum_i^p X_j \beta_j\right)$

Tabela 4.1 - Do Modelo Linear Clássico aos Modelos Lineares Generalizados

Assim, uma das características dos Modelos Lineares Generalizados é que μ_i é uma função do preditor linear η_i, sendo este uma combinação linear das p variáveis explicativas X_{i1}, …, X_{ip} na forma:

$$\mu_i = g^{-1}\left(\beta_1 X_{i1} \cdots \beta_{pX_{ip}}\right) \qquad (4.1)$$

Ou seja, tal como vimos em 3.2.3., θ é uma função composta dos elementos de β[34]:

$$\theta_i = b'^{-1}\left[g^{-1}\left(\beta_1 X_{i1} \cdots \beta_{pX_{ip}}\right)\right] \qquad (4.2)$$

Esta expressão mostra a forma como a distribuição de Y_i depende dos parâmetros β_1, …, β_p.

[33] Assim, a variância não é necessariamente constante, devendo apenas ser uma função da média.
[34] $\theta_i = (b')^{-1}(\mu_i)$

4.2. A MODELAÇÃO DOS DADOS

A modelação de dados através dos Modelos Lineares Generalizados segue cinco etapas:

1	**Análise preliminar dos dados e formulação do modelo**	Escolha da amostra a analisar; Escolha das componentes do Modelo.
2	**Ajustamento do modelo (ou modelos)**	Estimação dos parâmetros do Modelo; Análise da adequabilidade das estimativas obtidas; Realização de testes de ajustamento.
3	**Selecção e validação dos modelos**	Selecção do modelo mais adequado ao problema, incluíndo a análise de sub-modelos adequados; Análise de eventuais discordâncias entre os dados e os valores estimados.
4	**Ajustamento do modelo**	Correcção/ ajustamento de eventuais situações detectadas nas etapas anteriores.
5	**Interpretação dos resultados**	Comparação dos resultados produzidos pelo Modelo com a realidade existente e sua interpretação

Tabela 4.2 - Etapas da Modelação de Dados através dos Modelos Lineares Generalizados

Analisemos cada uma destas fases da modelação:

4.2.1. Análise Preliminar dos dados e Formulação do modelo

Nesta primeira fase, preparam-se os dados a modelar e fazem-se algumas análises estatísticas preliminares dos mesmos, que auxiliarão à escolha da amostra a utilizar e de algumas das componentes dos Modelos Lineares Generalizados.

4.2.1.1. A escolha da amostra a analisar

A primeira questão que se coloca é a da escolha das observações a modelar. Os princípios básicos a considerar nesta escolha são:

- A amostra utilizada ter um volume de informação considerado suficiente, e ser, tanto quanto possível, homogénea;
- Os dados utilizados devem ser credíveis;
- Os dados utilizados devem ser relevantes, ou seja, devem reflectir, tanto quanto possível, a realidade actual que se pretende modelar.

Para além disso, deverá ter-se em conta que os Modelos Lineares Generalizados utilizam a Estimação pelo Método da Máxima Verosimilhança, normalmente não adequado a amostras pequenas, como já referido.

4.2.1.2. A escolha das variáveis explicativas

Em algumas situações, como é o caso da tarifação, as variáveis explicativas têm vários níveis. A escolha das variáveis explicativas, bem como dos seus níveis, partirá sempre de uma análise empírica do problema em estudo. Devem também ser efectuadas análises estatísticas simples, como por exemplo, tabelas de frequência, para cada factor que permitirão verificar quais os factores que têm um efeito mais significativo no problema em estudo, bem como verificar se existem tendências que possam influenciar os resultados.

4.2.1.3. A escolha da função de ligação

Quanto à função de ligação, teoricamente, para cada observação pode ser utilizada uma função de ligação diferente, mas, na prática, tal normalmente não acontece.

Algumas das escolhas mais comuns para a Função de Ligação são:

	Identidade	Logaritmo	Logit	Log-Log Complementar	Inversa
$\eta = g(\mu)$	μ	$\ln(\mu)$	$\ln\left(\frac{\mu}{1-\mu}\right)$	$\ln(-\ln(1-\mu))$	$1/\mu$
$\mu = g^{-1}(\eta)$	η	e^{η}	$\frac{e^{\eta}}{1+e^{\eta}}$	$-\ln(1-\eta)$	$1/\eta$

Tabela 4.3 - Escolhas mais comuns para a função de ligação

Cada uma das distribuições da família exponencial tem associada a correspondente **função de ligação canónica**, que é tal que[35] $\theta = \eta$.

As funções de ligação canónica associada às mais comuns distribuições da família exponencial são:

[35] $\theta = (b')^{-1}(g^{-1}(\eta)) = \eta$

Normal	Poisson	Binomial	Gama	Gaussiana Inversa
μ	$\ln\mu$	$\ln\left(\frac{\mu}{1-\mu}\right)$	$1/\mu$	$1/\mu^2$

Tabela 4.4 - Funções de Ligação Canónica mais comuns

Utilizando a função de ligação canónica, obtemos sempre valores admissíveis para μ. Por exemplo, no caso da Distribuição de Poisson, a utilização da função de ligação logarítimica garante que μ tomará sempre valores positivos. No entanto, tal não significa que a função de ligação canónica seja sempre a escolha mais adequada, dependendo essa da situação em estudo e da forma como se pretendem modelar os dados. Nessa escolha, deverá ter-se em conta que, utilizando a função:

$\Rightarrow$ Identidade , $g(\mu_i) = \mu_i$ - temos, $E[Y_i] = g^{-1}(\eta_i) = \eta_i = \sum_{i,j} X_{ij}\beta_i$,

$\Rightarrow$ Inversa, $g(\mu_i) = \frac{1}{\mu_i}$ - donde, $E[Y_i] = g^{-1}(\eta_i) = \frac{1}{\eta_i} = \frac{1}{\sum_{i,j} X_{ij}\beta_i}$,

$\Rightarrow$ Logarítmica, $g(\mu_i) = \ln(\mu_i)$ - donde $E[Y_i] = g^{-1}(\eta_i) = \exp\{\eta_i\} = \exp\left\{\sum_{i,j} X_{ij}\beta_i\right\}$.

Ou seja, utilizando as funções de ligação Identidade e Inversa, modelamos os dados de forma **aditiva**. Note-se no entanto que, pelo facto de a Estimativa de Máxima Verosimilhança poder ser negativa para algum dos $\beta_i's$, é possível obtermos valores negativos para a média de Y, o que pode nem sempre ser adequado à situação prática em causa.

Utilizando a função de ligação logaritmo, **os dados são modelados de forma multiplicativa**. Para além disso, dado que $g^{-1}(.)$ é a função exponencial, tal permite obter apenas valores positivos para a média de Y.

4.2.1.4. A escolha da distribuição da variável resposta

Quanto à escolha da distribuição da variável resposta, deve ter-se em conta a natureza contínua ou discreta dos dados. De facto, as distribuições Gama ou Normal Inversa, são adequadas a dados de natureza contínua, enquanto que dados de natureza discreta são

normalmente modelados por uma distribuição de Poisson ou Binomial. Para além deste indicador, o tipo de situação em estudo também deverá ser considerado nesta escolha[36].

Pode ser ajustada uma distribuição aos dados com vista a identificar características invulgares ou problemas existentes nos dados e que devam ser eliminados antes da modelação. No entanto, a escolha da distribuição da variável resposta não tem que coincidir com a distribuição teórica ajustável aos dados. De facto esta escolha está relacionada com a estrutura de variância do problema em estudo, como será visto mais à frente.

4.2.2. Ajustamento do modelo (ou modelos)

4.2.2.1. Estimação de β

Dado um determinado conjunto de observações, e sendo definido um modelo, com as componentes acima, os valores do vector β são estimados através do Método da Máxima Verosimilhança, método utilizado não só devido às suas propriedades, como também devido ao facto de permitir aplicar testes de hipóteses sobre os parâmetros do modelo e aferir da qualidade do ajustamento.

Considerando o Modelo Linear Generalizado definido em 4.1., vem:

$$L(\beta) = \prod_{i=1}^{n} f(y_i \mid \theta_i, \phi, \omega_i) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi, \varpi_i) \right\} \qquad (4.3)$$

E a log-verosimilhança é dada por:

$$\boxed{\ln L(\beta) = l(\beta) = \sum_{i=1}^{n} \left[\frac{(y_i \theta_i - b(\theta_i))}{a(\phi)} + c(y_i, \phi, \varpi_i) \right] = \sum_{i=1}^{n} l_i(\beta)} \qquad (4.4)$$

Sendo $l_i(\beta)$ a contribuição de cada observação y_i para a verosimilhança.

A Estimativa de Máxima Verosilhança para β obtém-se através da equação de verosimilhança:

$$\boxed{\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(\beta)}{\partial \beta_j} = 0, \quad j = 1, \cdots, p.} \qquad (4.5)$$

[36] *McCulagah e Nelder* (1983) propõem distribuições para várias situações "tipo". Mais tarde, *Turkman e Silva* (2000), apresentam um resumo deste trabalho (Ver Anexo A)

Tome-se $\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{\partial l_i(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i(\mu_i)}{\partial \mu_i}\frac{\partial \mu_i(\eta_i)}{\partial \eta_i}\frac{\partial \eta_i(\beta)}{\partial \beta_j}$.

Recorde-se que, por (3.12), $b'(\theta_i) = \mu_i$, pelo que $\frac{\partial l_i(\theta_i)}{\partial(\theta_i)} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$.

Por outro lado, por (3.12) e (3.13), vem [37] que: $\frac{\partial(\mu_i)}{\partial \theta_i} = b''(\theta_i) = \frac{\mathrm{var}(Y_i)}{a(\phi)}$ e, portanto [38], $\frac{\partial \theta_i(\mu_i)}{\partial \mu_i} = [b''(\mu_i)]^{-1} = \frac{a(\phi)}{\mathrm{var}(Y_i)}$

Temos ainda que: $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

Assim: $\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)}\frac{a(\phi)}{\mathrm{var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}$, donde as **equações de verosimilhança para** β são:

$$\sum_{i=1}^{n}\frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \cdots, p. \qquad (4.6)$$

A função score é o vector p-dimensional $s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta)$, onde $s_i(\beta)$ é vector de componentes $\frac{\partial l_i(\beta)}{\partial \beta_j}$ definido acima.

O elemento genérico de ordem j **da função score é:**

$$\sum_{i=1}^{n}\frac{(y_i - \mu_i)\cdot x_{ij}}{\mathrm{var}(Y_i)}\cdot\frac{\partial \mu_i}{\partial \eta_i} \qquad (4.7)$$

Como já referido, a solução das equações de verosimilhança não corresponde necessariamente a um máximo global da função $S(\beta)$, embora em muitos modelos o máximo local e global coincidam e, em alguns casos, sejam únicos. Os modelos utilizados no âmbito deste trabalho utilizam, como se verá no Capítulo 5, as distribuições Poisson e Gama, com função de ligação logarítmica. Estes modelos têm uma Estimativa de Máxima Verosimilhança única - veja-se, por exemplo, *Turkman e Silva* (2000).

[37] Relembrando que, por (3.13), $\mathrm{var}(Y_i) = a(\phi)b''(\theta_i)$.

[38] Dado que, por (3.12) $\theta_i = b'^{-1}(\mu_i)$

No entanto, assumindo a existência e unicidade das Estimativas de Máxima Verosimilhança, as equações de verosimilhança não têm, em geral, uma solução analítica. Nas situações mais usuais, que envolvem muitos dados, é necessário recorrer a técnicas numéricas iterativas que permitem obter estimativas para β, tais como, por exemplo, o **Método de Scores de Fisher** e o Algoritmo de Newton Raphson[39]. Por conveniência computacional, o método utilizado neste trabalho será o primeiro.

O método de scores de Fisher parte de uma estimativa inicial $\hat{\beta}^{(0)}$, e através da relação definida em (4.8) calcula as sucessivas iteradas:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left[\mathrm{I}(\hat{\beta}^{(k)}\right]^{-1} s(\hat{\beta}^{(k)}) \qquad (4.8)$$

em que $I(\beta) = E\left[-\frac{\partial s(\beta)}{\partial \beta}\right]$ é a matriz de covariância da função score, conhecida como a matriz de informação de Fisher.

4.2.2.2. Estimação do parâmetro de escala ϕ

Em algumas situações, tais como a Distribuição de Poisson, o parâmetro de escala é conhecido, como veremos mais à frente. No entanto, em geral, este parâmetro não é conhecido à priori, e tem que ser estimado.

A estimação do parâmetro de escala não é necessária para a obtenção do vector β[40], mas sim para determinar algumas estatísticas, tais como o erro, como se verá na subsecção 4.2.3..

Para estimar ϕ pode ser também utilizado o Método da Máxima Verosimilhança, que no entanto não permite a obtenção de uma fórmula explícita para ϕ, e pode ser uma alternativa mais lenta. Em geral, utilizam-se estimadores para ϕ:

- O estimador de momentos ou Estatística χ^2 de Pearson, definido como :

[39] Veja-se, por exemplo, *Andersen, Feldblum, Modlin, Schirmacher, Schirmacher, and Thandi* (2004)

[40] Dado que a Equação de Verosimilhança não depende de ϕ

$$\hat{\phi} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{V(\mu_i)} \qquad (4.9)$$

- O estimador baseado na Estatística de Pearson Generalizada[41], um estimador consistente e assintoticamente centrado:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{\varpi_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \qquad (4.10)$$

- O Estimador "*Deviance*", definido como:

$$\hat{\phi} = \frac{D}{n-p} \qquad (4.11)$$

sendo D a "*Deviance*", que será definida mais à frente.

4.2.2.3. O termo Offset

Por vezes, o efeito da variável explicativa é conhecido, e, em vez de estimarmos os parâmetros β, é aconselhável incluir no modelo informação sobre essa variável, como sendo um efeito conhecido. Tal pode ser alcançado, através da introdução de um termo Offset, ξ, na definição do preditor linear, ou seja, $\eta = X\beta + \xi$, donde: $E[Y] = \mu = g^{-1}(\eta) = g^{-1}(X\beta + \xi)$.

4.2.2.4. Testes de hipóteses sobre β

A maior parte dos testes de hipóteses sobre o parâmetro β, podem ser formulados na forma $H_0 : C\beta = \varepsilon \; vs \; H_1 : C\beta \neq \varepsilon$, onde C é uma matriz q x p, com q ≤ p e ε é um vector de dimensão q previamente especificado. São casos particulares desta forma a hipótese de nulidade de uma componente do vector parâmetro.

Existem três tipos de estatísticas para testar hipóteses da forma geral acima indicada:

[41] $\sum \frac{\varpi_i (Y_i - \hat{\mu}_i)^2}{V(\hat{\mu})}$. Este é o estimador utilizado pelo software estatístico usado na componente prática deste trabalho.

- **Estatística de Wald** – Em geral, é mais utilizada para testar hipóteses nulas sobre componentes individuais, embora em algumas situações possa ser utilizada para testar hipóteses nulas sobre um subvector. A Estatística de Teste é:

$$(\hat{\beta}-\beta)^T \mathrm{I}(\beta)(\hat{\beta}-\beta) \overset{a}{\sim} \chi^2_p \qquad (4.12)$$

- **Estatística de Wilks ou Estatística de razão de verosimilhanças –** é normalmente utilizada para testar modelos que estão encaixados, ou seja, modelos em que um é sub-modelo do outro. A Estatística de Teste é:

$$\Lambda = -2\ln\frac{\max_{H_0} L(\beta)}{\max_{H_0 \cup H_1} L(\beta)} = -2\left\{l(\tilde{\beta}) - l(\hat{\beta})\right\} \qquad (4.13)$$

onde $\tilde{\beta}$, o estimador de máxima verosimilhança restrito, é o valor de β que maximiza a verosimilhança sujeito ás restrições impostas pela hipótese $C\beta = \varepsilon$. Sob certas condições de regularidade[42], estabelecidas pelo Teorema de Wilks, $\Lambda \overset{a}{\sim} \chi^2_q$

- **Estatística de Rao ou Estatística Score –** Útil em situações em que já se calculou um estimador restrito para β, com a vantagem, em relação a Estatística de razão de verosimilhanças, de não requerer o cálculo do estimador não restrito. Esta estatística mede a distância entre $S(\tilde{\beta})$ e 0. A Estatística de Teste é:

$$\Psi = \left[S(\tilde{\beta})\right]^T \mathrm{I}(\tilde{\beta}) S(\tilde{\beta}) \overset{a}{\sim} \chi^2_q \qquad (4.14)$$

4.2.3. Selecção e validação dos modelos

Quando se estuda um problema em que existem muitas possíveis variáveis explicativas a considerar, uma das análises a efectuar é sobre qual o modelo mais adequado, ou seja, o que tem menos variáveis explicativas e que ainda assim permite uma boa interpretação do problema em estudo e que se ajusta bem aos dados.

Consideremos o modelo Saturado (ou completo) e o modelo Corrente. O primeiro trata-se de um modelo que representa os próprios dados, e portanto, um modelo teórico[43]. O Modelo Corrente é um modelo com q parâmetros, que, embora não tenha o maior número de

[42] *Turkman e Silva* (2000)
[43] Na medida em que não permite simplificar o problema nem visualizar situações particulares transmitidas pelos dados.

parâmetros possível, tem um número de parâmetros que permite modelar a situação sem esconder características importantes dos dados.

4.2.3.1. Qualidade do ajustamento

As medidas mais utilizadas para aferir da bondade do ajustamento são a *Deviance* e a Estatística de Pearson Generalizada.

i. A *Deviance*

A ***Deviance***, em termos gerais, é uma medida de quanto os valores ajustados diferem das observações.

Designemos o modelo em estudo por M, e por $\hat{\mu}_i$, a estimativa de máxima verosimilhança para μ_i nesse modelo. O modelo saturado, que designaremos por S, permitir-nos-á aferir da qualidade de ajustamento de M, através da introdução de uma medida da distância dos valores ajustados com esse modelo e os valores observados, ou seja, a distância entre $\hat{\mu}_i$ e y. Se compararmos M com S, através da Estatística de Wilks, já definida atrás, obtemos as medidas ***Deviance***, denotada por $D(y,\hat{\mu})$ e ***Deviance* Reduzida** ou ***Deviance* à escala**, denotada por $D^*(y,\hat{\mu})$.

Tendo em conta que θ_i é função de μ_i, explicitemos essa relação, substituíndo, na Log-Verosimilhança, $\hat{\theta}_i$ por $q(\hat{\mu}_i)$. Dado que, no modelo S, temos $\mu_i = y_i$, obtemos então:

$$D^*(y,\hat{\mu}) = -2(l_M(\hat{\beta}_M) - l_S(\hat{\beta}_S))$$

$$D^*(y,\hat{\mu}) = -2\sum_i \frac{\varpi_i}{\phi}\cdot\{[y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i))] - [y_i q(y_i) - b(q(y_i))]\} = \frac{D(y;\hat{\mu})}{\phi} \qquad (4.15)$$

sendo que esta última pode ser decomposta na soma de parcelas que medem a diferença entre os logaritmos das verosimilhanças observada e ajustada para cada observação, ou seja:

$$D(y;\hat{\mu}) = \sum_i 2\cdot\varpi_i\cdot\{y_i(q(y_i) - q(\hat{\mu}_i)) - b(q(y_i)) + b(q(\hat{\mu}_i))\} = \sum_i d_i \qquad (4.16)$$

Assim, a *Deviance* é uma medida da discrepância entre as duas log-verosimilhanças.

Considere-se o modelo nulo, ou seja, o modelo mais simples, que possui apenas um parâmetro. A *Deviance* é sempre maior ou igual a zero e decresce à medida que variáveis explicativas são adicionadas ao modelo nulo, sendo, obviamente, zero para o modelo saturado.

Em geral, nem a distribuição exacta nem a distribuição assintótica da *Deviance* são conhecidas. Assim, a *Deviance* pode ser utilizada para avaliar a adequabilidade de um modelo, mas deve ser considerada apenas como um guia. Na prática, normalmente compara-se o valor observado da *Deviance* Reduzida com o valor crítico de uma distribuição χ^2_{n-p}, onde p é o número de parâmetros do modelo, considerando o modelo não adequado quando o valor esperado da *Deviance* Reduzida for superior a $\chi^2_{n-p,\alpha}$.

ii. A Estatística de Pearson Generalizada

Tal como já definido em 4.2.2.2., a Estatística de Pearson Generalizada é:

$$X^2 = \sum_i \frac{\varpi_i (Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \qquad (4.17)$$

Normalmente, testa-se a adequabilidade de um modelo, através desta estatística, comparando o valor observado com o quantil de probabilidade $\chi^2_{n-p;\alpha}$. No entanto, tal como acontece com a Deviance, esta aproximação pode ser má em certos modelos, mesmo para grandes amostras, pelo que esta medida deverá também ser considerada apenas como um guia da adequabilidade do modelo.

4.2.3.2. Selecção dos Modelos

Normalmente, não existe um único modelo adequado, pelo que podem ser utilizadas algumas ferramentas para nos auxiliar na escolha do modelo a utilizar:

Supondo dois modelos intermédios M_1 e M_2, com M_2 encaixado[44] em M_1, ou seja dois modelos do mesmo tipo, mas em que M_2 tem menos parâmetros que M_1. Sendo $D(y, \hat{\mu}_j)$ a Deviance para o modelo M_j, com j=1,2, então a Estatística da Razão de Verosimilhanças pode escrever-se como[45]:

[44] Turkman e Silva (2000)
[45] Tendo em conta os resultados descritos no ponto 4.2.3.1

$$-2(l_{M_2}(\hat{\beta}_2) - l_{M_1}(\hat{\beta}_1)) = \frac{D(y;\hat{\mu}_2) - D(y;\hat{\mu}_1)}{\phi} \overset{a}{\sim} \chi^2_{p_1-p_2} \qquad (4.18)$$

onde p_j representa a dimensão do vector β para o modelo M_j. A *Deviance* é aditiva para modelos encaixados.

Assim sendo, a comparação de modelos encaixados pode ser feita pela diferença da *Deviance* de cada modelo, o que permite aferir se a inclusão de uma variável explicativa no modelo o melhora de forma significativa. Ou seja, caso a diferença entre a *Deviance* M_1 e M_2 não seja significativa, então podemos optar pelo modelo M_2, dado este ser mais simples que M_1.

Podem ainda ser efectuados Testes de Hipóteses, baseados nas Estatísticas descritas em 4.2.2.4., bem como ser utilizado o critério de informação de Akaike, baseado na função log-verosimilhança, com a introdução de um factor de correcção como forma de penalização da complexidade do modelo. Estas análises podem ser consultadas em *Turkman e Silva* (2000).

4.2.3.3. Análise de resíduos e do erro como análise da qualidade de ajustamento do modelo

O **Erro Padrão** corresponde à diagonal da matriz de covariâncias $-H^{-1}$, em que H, designada de matriz Hessiana, tem como elementos a segunda derivada dos elementos da matriz de log-verosimilhança. O erro padrão de um determinado parâmetro pode ser visto como um indicador da "velocidade" com que uma alteração no parâmetro faz com que o estimador de log-verosimilhança se afaste do máximo.

A análise de resíduos permite avaliar a qualidade do ajustamento, mas também permite ajudar a identificar observações mal ajustadas, ou seja, que não são bem explicadas pelo modelo. Os resíduos permitem ainda verificar os pressupostos assumidos na formulação do modelo, devido ao facto de estes serem normalmente independentes, não relacionados com as variáveis explicativas e com uma distribuição aproximadamente Normal, com média zero e variância constante.

Um Resíduo deve exprimir a discrepância entre o valor observado e o valor ajustado pelo modelo para o seu valor médio.

A definição "clássica" de Resíduos, ou seja, os **Resíduos Simples** é a diferença entre os valores observados e o valor esperado previsto pelo Modelo. Para além destes, existem

várias definições de resíduos, sendo as mais utilizadas o Resíduo de Pearson, a *Deviance* Residual e a *Deviance* Residual Padronizada[46].

O **Resíduo de Pearson** corresponde à contribuição de cada observação para o cálculo da Estatística de Pearson Generalizada, ou seja:

$$R_i^P = \frac{(y_i - \hat{\mu}_i)\varpi_i}{\sqrt{\hat{\phi}V(\hat{\mu}_i)}} \qquad (4.19)$$

A **Deviance Residual** é a raiz quadrada da contribuição de cada observação para a Deviance Total[47], d_i, tal como definida por (4.16), multiplicada por 1 ou -1, consoante cada observação seja maior ou menor que cada valor ajustado, ou seja, é dada pela expressão:

$$R_i^D = \delta_i \sqrt{d_i}, \text{ em que } \delta_i = Sinal(y_i - \hat{\mu}_i) \qquad (4.20)$$

Geralmente, a *Deviance* Residual apresenta uma distribuição mais próxima da Normal que os Resíduos Simples, já que o cálculo da *Deviance* Residual corrige a Distorção[48] das distribuições. Quando trabalhamos com distribuições contínuas, qualquer desvio grande da distribuição Normal é uma boa indicação de que as assumpções da distribuição não estão correctas.

Definidos os resíduos, a questão que se coloca é como analisá-los e utilizá-los para aferir da bondade do ajustamento produzido pelo Modelo.

Uma das análises de resíduos mais utilizada é a sua representação gráfica contra os seus valores esperados. Os pontos deverão concentrar-se sobre ou perto de uma linha recta representando a normalidade e qualquer desvio sistemático indicia um afastamento da Normalidade, permitindo detectar anomalias no modelo.

A representação gráfica dos resíduos contra o estimador de η ou contra transformações adequadas do valor estimado de μ[49] é uma análise que permite detectar anomalias como a escolha errada da função de ligação e a escolha errada da escala. Caso não existam anomalias, os resíduos devem estar distribuídos em torno de zero com uma amplitude constante para diferentes valores de $\hat{\mu}$.

[46] Outras definições, como o Resíduo Padronizado de Pearson, a Deviance Residual Padronizada e o Resíduo de Anscombe, podem ser encontradas em *Turkman e Silva* (2000)
[47] Ou seja, uma medida da distância entre as observações e os estimadores.
[48] A Distorção caracteriza o grau de assimetria de uma distribuição em redor do seu ponto médio.
[49] *McCullagh e Nelder* (1989) sugerem transformações para alguns modelos.

4.2.4. Re-Ajustamento do modelo

Em função da análise anterior, poderá ser necessário reajustar o modelo, ao nível da escolha das componentes do modelo. Poderá também ser necessário introduzir restrições no modelo, decorrentes, por exemplo, de imposições legais ou comerciais, no caso da tarifação.

4.2.5. Interpretação dos resultados

Os resultados produzidos pelo Modelo Linear Generalizado podem ser diferentes da realidade já existente. Assim, a análise dos resultados deverá ser feita de forma crítica, tendo em conta a realidade concreta em estudo, bem como a aplicabilidade dos mesmos resultados. Ao comparar as estimativas do modelo com os valores reais, pode verificar-se um desequilíbrio ao nível dos valores de cada variável explicativa individualmente. No entanto, caso as estimativas obtidas sejam mais gravosas que a realidade nuns casos, e menos gravosas noutros, tal deverá ser tido em conta na análise final, dado que, em termos globais, a situação pode estar equilibrada.

4.3. PORQUE UTILIZAR OS MODELOS LINEARES GENERALIZADOS?

Os Modelos Lineares Generalizados, como vimos neste capítulo, são aplicáveis a problemas em que se pretende estudar o efeito que determinadas variáveis têm nas observações, pelo que são aplicáveis na Tarifação, problema em que se pretende modelar a sinistralidade e estudar os efeitos que cada factor e cada um dos seus níveis de tarifação têm nessa mesma sinistralidade.

Uma das vantagens destes modelos é que têm a flexibilidade de se poder especificar uma função de ligação e a distribuição que representa as observações - embora esta deva pertencer à família de distribuições exponencial - o que aumenta o rigor do ajustamento. Para além disso, os Modelos Lineares Generalizados permitem o cálculo dos resíduos e de medidas como a *Deviance* e a Estatística de Pearson Generalizada, que nos dão informação sobre a bondade do ajustamento do modelo.

Por outro lado, a sua estrutura comum e generalizada, facilita a sua compreensão e a sua implementação informática, o que se tem verificado com a expansão de software informático

que permite a modelação de dados através destes modelos, e tem facilitado a utilização deste modelos.

Por estas razões, foi esta a via escolhida para o estudo da Tarifa, nesta dissertação. Veremos, de seguida, as particularidades da aplicação dos Modelos Lineares Generalizados na Tarifação.

5. OS MODELOS LINEARES GENERALIZADOS APLICADOS À TARIFAÇÃO – ANÁLISE PRELIMINAR DOS DADOS E FORMULAÇÃO DO MODELO

Neste capítulo, aborda-se a primeira fase da Modelação de Dados, no que se refere ao caso particular da tarifação. Começa por ser abordada a análise preliminar dos dados, ao nível da selecção da amostra a modelar. Aborda-se de seguida a escolha das componentes dos Modelos Lineares Generalizados.

Comecemos por lembrar que o Prémio Puro é obtido através do valor esperado das Indemnizações (S). Uma alternativa possível é modelar directamente $E(S)$ e obter directamente o prémio puro, utilizando a distribuição Tweedie[50], distribuição da família Exponencial, e que, genericamente, é uma distribuição composta, que permite incorporar a distribuição da incidência de sinistros e da gravidade do seu custo. No entanto, sabemos que a Frequência e a Severidade de sinistros normalmente seguem distribuições diferentes. Para além disso, modelar a Frequência e a Severidade de sinistros separadamente proporciona um entendimento mais detalhado sobre a forma como cada um dos factores afectam o valor esperado das Indemnizações e sobre a volatilidade dos sinistros. Permite ainda detectar e eliminar mais facilmente efeitos aleatórios de um determinado elemento. Por estas razões, no âmbito deste trabalho, o Prémio Puro será obtido pela modelação separada da Frequência e da Severidade de sinistros.

5.1. A ESCOLHA DA AMOSTRA A ANALISAR

5.1.1. Período a analisar

Tendo em conta os princípios já enunciados em 4.2.1.1., por um lado, deve pesar-se a relevância da utilização de dados menos recentes, avaliando se as projecções obtidas por esses dados reflectirão o futuro que estamos a modelar. No entanto, embora os dados mais recentes sejam, geralmente, mais representativos da realidade, não é aconselhável utilizar dados baseados apenas num ano de sinistralidade, já que os mesmos têm implícito um maior grau de incerteza. Este grau de incerteza provém do facto de poderem ainda existir sinistros tardios desconhecidos e de, habitualmente, existir um maior número de sinistros não encerrados, cujos valores em reserva têm maior probabilidade de sofrerem alterações.

[50] Veja-se, por exemplo, *Andersen, Feldblum, Modlin, Schirmacher, Schirmacher, e Thandi* (2004)

Assim, geralmente, utilizam-se no mínimo três anos de experiência, período que pode ser estendido, tendo naturalmente em conta a realidade específica e a dimensão de cada seguradora.

Aliás, a realidade da seguradora é sempre um factor essencial a ter em conta nesta escolha. Deve considerar-se, por exemplo, o efeito que eventuais mudanças na política de subscrição da companhia e na política de regularização de sinistros possam ter tido nos dados seleccionados. Por exemplo, caso uma determinada cobertura tenha deixado de ser comercializada, a mesma deverá ser retirada da análise.

A sinistralidade de um determinado ano pode ser influenciada por factores extraordinários e que não se prevê que se repitam, tais como por exemplo, factores climatéricos anormalmente adversos. Esse ano pode ser retirado da análise, ou caso se entenda que deve ser mantido na análise[51], os dados podem ser ajustados por aplicação de um factor subjectivo que "corrija" a sinistralidade, de forma a reflectir melhor a realidade. Este factor de correcção pode ter em conta informação de mercado, se disponível, bem como o histórico da companhia, no que se refere às tendências registadas nos anos não atípicos.

5.1.2. Custos com Sinistros

O período seleccionado pode incluir sinistros pendentes que, por essa razão, incluem ainda montantes em reserva. Estes montantes deverão ser considerados na análise, dado serem montantes que a companhia espera vir a pagar e que, por essa razão, influenciam a projecção da Severidade de sinistros.

Em algumas situações, os montantes em reserva podem estar sub-estimados ou até sobre-estimados. Por outro lado, os sinistros já encerrados podem ser reabertos. Existem também sinistros ainda não conhecidos, mas que podem vir a ser declarados, os sinistros tardios. Estas questões podem ser consideradas na análise, incluídos no cálculo da carga de segurança, abordagem que será seguida neste trabalho. Uma alternativa seria incluir esta carga directamente na modelação, distribuindo o montante referente a esta carga por todos os sinistros a analisar.

Existem ainda sinistros de custo igual a zero, sinistros declarados à seguradora, mas encerrados sem despesas, normalmente por o evento participado não estar coberto pelo contrato de seguro. A inclusão destes sinistros pode contribuir para uma subestimação da Severidade de sinistros, pelo que os mesmos deverão ser eliminados do estudo. Dado que

[51] Por exemplo, por ser um ano bastante recente.

estes sinistros têm custos administrativos para a seguradora, a carga de segurança deverá ter estes custos em consideração.

Por outro lado, no Ramo Responsabilidade Civil do ramo Automóvel, podem existir sinistros com custos negativos. Trata-se de sinistros regularizados no âmbito da Convenção "Indemnização Directa ao Segurado"[52]. Estes custos não devem ser ignorados, por se tratarem de montantes efectivamente recebidos pelas seguradoras. Neste trabalho, dado que o modelo seguido para a modelação da severidade de sinistros pressupõe que todos os dados são positivos, como veremos mais à frente, estes sinistros foram eliminados da base de dados, mas o seu custo foi distribuído pelos restantes sinistros. Ao efectuar esta simplificação, poderemos subestimar a projecção da frequência de sinistros, pelo que tal deverá ser tido em conta no cálculo da carga de segurança.

Há ainda que ter em conta que a amostra estudada deve ser, tanto quanto possível, constituída por dados homogéneos, pelo que, em princípio, sem prejuízo da modelação conjunta dos dados, devem ser estudadas separadamente tipologias de sinistros diferentes. De facto, tipos de sinistros diferentes podem afectar de forma diferente a Frequência e a Severidade de sinistros, bem como podem afectar de forma diferente os factores de tarifação. Por exemplo, no ramo Responsabilidade Civil Automóvel, é comum que os sinistros de Danos Materiais (genericamente, os danos causados aos veículos), sejam sinistros de maior frequência que os de Danos Corporais (genericamente, danos físicos ou mentais causados aos terceiros), mas com um Custo Médio inferior. Assim, a modelação dos sinistros por Tipo de Dano permite captar a volatilidade associada a cada um.

5.1.3. Grandes Sinistros

Na maioria das carteiras de sinistros, ocorrem alguns sinistros de natureza excepcional e de valor elevado. O custo médio destes sinistros é altamente variável e, devido ao facto de ser substancialmente mais elevado que o dos restantes, pode condicionar a fiabilidade e adequabilidade das estimativas obtidas. Assim, normalmente, é conveniente separar os grandes sinistros dos restantes. Desta forma, a modelação desses sinistros produz resultados mais estáveis que a modelação do conjunto dos sinistros. Utilizando este método

[52] Trata-se de uma convenção gerida pela Associação Portuguesa de Seguradores, que abrange, genericamente, os sinistros em que um dos condutores assume a culpa do sinistro. Nestas situações, cada seguradora indemniza directamente o seu segurado, mesmo que não tenha sido este o causador do sinistro. Posteriormente, a Associação Portuguesa de Seguradores organiza a regularização dos custos com sinistros entre as várias seguradoras envolvidas. Esta regularização não é feita caso a caso, mas sim através do custo médio de vários sinistros, pelo que, num determinado sinistro, a seguradora pode receber mais da congénere do que pagou ao seu segurado.

é naturalmente necessário adicionar ao prémio de risco uma margem adicional para fazer face aos custos futuros com grandes sinistros.

Em alguns ramos, a ocorrência de grandes sinistros pode ser considerada um acontecimento ocasional, que acontece apenas de n em n anos, pelo que poderá optar-se simplesmente por excluir os grandes sinistros do estudo, e adicionar 1/n do custo com sinistros graves ao prémio de risco anual, se algum já foi conhecido[53]. Caso ainda não haja sinistros conhecidos, deverá adicionar-se uma margem prudente[54] para fazer face a eventuais sinistros futuros.

No entanto, noutros ramos, como é o caso do ramo em estudo, a ocorrência destes sinistros é mais frequente, ocorrendo normalmente todos os anos. Assim, neste trabalho, optou-se por truncar o custo destes sinistros, mantendo-os na modelação da frequência de sinistros e mantendo também parte do seu custo na modelação da Severidade de sinistros. Assim, definido o valor a partir do qual se considera que se trata de um grande sinistro (Outlier)[55], considera-se, na modelação, esse montante como custo máximo.

Obtidas as estimativas para a Severidade de sinistros, a margem de risco a adicionar, que designaremos por θ_{GS}, poderá ser obtida pelo rácio[56]:

$$\theta_{GS} = \frac{Custo\ Total\ com\ Sinistros\ (Grandes + Pequenos)}{Custo\ Total\ com\ Pequenos\ Sinistros} - 1 \qquad (5.1)$$

Com esta abordagem, assume-se que a ocorrência de grandes sinistros não depende dos factores de tarifação (assumpção, em geral, aceite). No entanto, poderá ser efectuada uma análise preliminar aos dados, para verificar se na carteira em causa existe alguma tendência para a ocorrência de grandes sinistros em determinados níveis dos factores de tarifação. Nesse caso, aos níveis dos factores de tarifação que apresentem uma tendência para a ocorrência de grandes sinistros, deve ser aplicada uma margem de segurança mais elevada que aos restantes. Nesta situação, *Murphy, Brockman e Lee* (2000), propõem o ajustamento de um Modelo Linear Generalizado à modelação dos grandes sinistros. O modelo a ajustar tem como componentes o número de "Grandes Sinistros" como variável resposta, e os factores de tarifação como variáveis explicativas. Assume-se para a variável Y a distribuição Binomial, e o número de sinistros total como o número de provas. O resultado desta

[53] *Portugal* (2007)
[54] Esta margem pode ser calculada com base em informação de mercado, se disponível.
[55] De acordo com o perfil de sinistros da carteira estudada.
[56] Veja-se *Murphy, Brockman e Lee* (2000)

modelação dará uma indicação de quais os factores aos quais devem ser aplicados ajustamentos mais elevados, como carga de segurança.

5.1.4. A inflação

A inflação influencia os custos com sinistros, pelo que deve ser considerada na modelação da severidade. Assim, os montantes já pagos devem ser actualizados a preços actuais. Quanto aos montantes em reserva, os mesmos deverão já reflectir o custo futuro esperado a preços actuais, pelo que não é necessário os mesmos serem actualizados[57].

O valor da inflação considerado pode ser diferente consoante o ramo e/ ou a tipologia de sinistros em estudo. No ramo Responsabilidade Civil Automóvel, a inflação reflecte-se nos custos com sinistros, no que se refere aos aumentos de custos das peças, da mão de obra e dos veículos novos, no caso dos Danos Materiais e nos aumentos das despesas hospitalares e da "inflação judicial"[58], no caso dos Danos Corporais.

5.2. ESCOLHA DAS VARIÁVEIS EXPLICATIVAS E SEUS NÍVEIS

Os princípios a ter como referência nesta escolha são de que os factores de tarifação devem, tanto quanto possível, fornecer uma boa aproximação ao risco e, por outro lado, serem mensuráveis, de utilização prática e aceites, tanto a nível comercial como legislativo. A estrutura tarifária já existente pode servir de base inicial à escolha tanto dos factores como da classificação dos vários níveis de cada factor. Podem ainda ser considerados níveis diferentes dos já existentes e factores não considerados na estrutura tarifária vigente. Para seleccionar estes novos factores, pode ter-se como referência informação de mercado (se disponível) ou informação disponível na base de dados e não utilizada no método de tarifação actual[59].

A análise estatística referida em 4.2.1.2., no caso da tarifação, pode ser a obtenção, para cada factor e nível de tarifação, de indicadores como a frequência de sinistros, o custo médio, o prémio puro e o Loss Ratio. Esta análise dar-nos-á não só uma primeira visão global de cada factor, como nos permitirá analisar se os sinistros se concentram apenas num dos níveis ou em vários, o que poderá ser útil na interpretação dos resultados obtidos na modelação. Posteriormente, tendo em conta esta análise estatística, bem como as análises

[57] Caso se entenda que existe o risco de desenvolvimento adverso dos valores em reserva, tal deverá ser tido em conta no cálculo da margem de segurança, tal como referido em 5.1.2.
[58] Por "inflação judicial", entenda-se as mudanças na jurisprudência, que podem aumentar os valores das indemnizações às vítimas.
[59] Desde que a informação disponível permita respeitar os princípios referidos em 4.2.1.1.

de validação do modelo, os níveis com uma reduzida exposição poderão ser agrupados noutro nível.

Na prática, dado que se utilizam modelos com muitos factores, e estes com muitos níveis, é útil considerar Níveis Base, que têm em comum um "Termo de Intercepção" [60], que se aplica a todos os níveis. Ou seja, selecciona-se o nível de cada factor que corresponde à situação "standard", pelo que não é necessário, para esse nível, estimar os parâmetros β_i.

No caso do Ramo Responsabilidade Civil, as tarifas incluem um sistema de Bónus Malus, que se baseia no histórico da frequência de sinistralidade da apólice. Sendo esta uma variável existente na Tarifa, mas cujo objectivo é o de tarifar o risco à posteriori, coloca-se a questão sobre se a mesma deve ou não ser incluída na modelação.

Alguns autores defendem que o estudo do sistema Bónus Malus é, essencialmente, um problema estocástico. De facto, alguns trabalhos já foram desenvolvidos nesta base [61], procurando não só analisar a percentagem de desconto ou agravamento de cada classe, mas também as regras de transição entre classes.

No contexto dos Modelos Lineares Generalizados, normalmente nem a estrutura nem as regras de transição do sistema de Bónus Malus são modeladas. Existem duas abordagens mais comummente seguidas: a escala de Bónus Malus é fixada, e a percentagem de desconto ou agravamento é incluída na modelação, ou, alternativamente, a percentagem de desconto ou agravamento é também fixada, sendo a mesma incluída no preditor linear para cada classe.

5.3. ESCOLHA DA FUNÇÃO DE LIGAÇÃO – MODELO DE CÁLCULO DO PRÉMIO

No caso da tarifação da cobertura Responsabilidade Civil Automóvel, o modelo mais comummente utilizado é o modelo multiplicativo. Por um lado, este permite a construção da tarifa em torno de um prémio base, facilmente actualizável periodicamente, se necessário, nomeadamente em função de alterações externas, tais como, por exemplo, o aumento da inflação. Por outro lado, a modelação dos dados através do modelo aditivo pode originar a obtenção de estimativas negativas, o que não acontece no modelo multiplicativo. Para além disso, empiricamente, a opção por um modelo multiplicativo também parece ser a mais

[60] Optou-se por esta tradução para "Intercept Term" (*Andersen, Feldblum, Modlin, Schirmacher, Schirmacher, e Thandi (2004)).*

[61] Não sendo esse o objecto do presente trabalho, alguns trabalhos neste âmbito podem ser consultados, por exemplo, em www.actuaries.org - International Actuarial Association (IAA) e www.casact.org - Casualty Actuarial Society.

adequada à modelação dos sinistros automóvel. Por exemplo, consideremos a situação proposta por *Brockman e Wright (1992):*

Idade do Condutor	30	30	17	17
Classe do Veículo	1	6	1	6
Frequência de Sinistros	5%	10%	20%	**40% ou 25%?**

No contexto do risco no ramo automóvel, parece ser mais plausível que a Frequência de sinistros do veículo de classe 6 seja o dobro da Frequência de sinistros de um veículo da classe 1, independentemente da idade do condutor, ou seja, de 40%, tal como seria obtido pelo modelo multiplicativo, e não de 25%, como obteríamos aplicando um modelo aditivo. O mesmo raciocínio pode ser seguido para outros factores de tarifação, bem como para a Severidade de sinistros.

O modelo multiplicativo tem as vantagens práticas já acima referidas e é um modelo simples mas ainda assim fiável, pelo que é o modelo adoptado neste trabalho, tanto no que se refere à modelação da Frequência de sinistros, como à modelação da Severidade de sinistros. Assim, a função de ligação a utilizar é a logarítmica, tal como exposto em 4.2.1.3.

Assim, sendo Y_i a variável resposta para o problema em questão, tal como definido na subsecção 4.1., a formulação do modelo é dada por:

$$\eta_i = \ln(\mu_i) \qquad (5.2)$$

e

$$E[Y_i] = g^{-1}(\eta_i) = \exp\left\{\sum_{i,j} X_{ij}\beta_i\right\} \qquad (5.3)$$

5.4. ESCOLHA DA DISTRIBUIÇÃO DA VARIÁVEL RESPOSTA NA MODELAÇÃO DA FREQUÊNCIA DE SINISTROS

Segundo exposto em *Nelder e McCullagh (1989)*, o modelo de Poisson é adequado para modelar dados do tipo contagens sem valores máximos. É também geralmente aceite que a distribuição de Poisson é adequada para modelar o número de sinistros. Assim, intuitivamente, o modelo de Poisson parece ser a primeira escolha para modelar a Frequência por factor de tarifação. De facto é este o modelo proposto por vários autores e cuja adequabilidade tem sido confirmada por várias aplicações práticas.

Comecemos então por especificar a componente aleatória do modelo: dado que trabalharemos com dados em forma de contagens, nomeadamente o número de sinistros, para modelar a Frequência de sinistros, assumimos que Y tem uma distribuição de Poisson, ou seja:

$Y \sim P(\lambda)$, com função de probabilidade $f(y \mid \lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}$, $y = 0,1,\cdots \quad e\ \lambda > 0$ *(5.4)*

Ora:

$$f(y \mid \lambda) = \exp\{\ln(e^{-\lambda}) + \ln(\lambda^{y}) - \ln(y!)\} = \exp\{y.\ln(\lambda) - \lambda - \ln(y!)\}$$

Fazendo $\lambda = e^{\theta}$ (donde $\theta = \ln(\lambda) = \ln(\mu)$), vem:

$$f(y \mid \lambda) = \exp\{y\theta - e^{\theta} - \ln(y!)\} = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) = f(y \mid \theta,\phi), \text{ com } a(\phi) = 1 \text{ e } \phi = 1.$$

Ou seja, a distribuição de Poisson é uma distribuição da família exponencial, em que:

Distribuição de **Poisson** como membro da Família Exponencial	θ	$b(\theta)$	ϕ	$a(\phi)$	$c(y,\phi)$
	$\ln(\lambda)$	e^{θ}	1	1	$-\ln(y!)$

Tabela 5.1 - A Distribuição de Poisson na Família Exponencial

Temos que:

$$E[Y] = \mu = b'(\theta) = e^{\theta},\ \operatorname{var}(Y) = b''(\theta)\cdot a(\phi) = e^{\theta} = \lambda \text{ e } V(\mu) = e^{\theta}$$

Tomando a função de ligação logarítmica [62], o modelo fica especificado por $\eta_i = \ln(\mu_i) + \varepsilon = \ln(e^{\theta_i}) = \theta_i$, ou seja, $\eta = \theta$ (donde, ln é a função de ligação canónica do Modelo de Poisson).

No problema em questão, as observações de Y agrupam-se em i factores de Tarifação e cada observação pode reportar-se a períodos de exposição da apólice diferentes, ou seja, temos:

- A variável resposta Y_{ij}, a Frequência de sinistros no Factor de Tarifação i, Nível j

[62] Dado que o modelo de cálculo do prémio escolhido é o modelo multiplicativo.

- O Número de Sinistros do Factor de Tarifação i, Nível j , n_{ij}
- A Exposição do Factor de Tarifação i, Nível j, e_{ij}

Suponhamos que os factores de tarifação foram seleccionados de forma a que exista homogeneidade em cada factor/nível, ou seja, que em todas as unidades de exposição do Factor de Tarifação i, Nível j exista a mesma probabilidade de ocorrência de 1, 2, 3 ou mais sinistros. Suponha-se ainda que o número de sinistros de cada unidade de exposição segue uma distribuição de Poisson, em que $E\left[n_{ij}\right] = \mathrm{var}\left[n_{ij}\right] = m_{ij}$, ou seja, a variância é igual á média.

Temos que **:**

$$\boxed{Y_{ij} = \frac{1}{e_{ij}} \cdot \sum_{i,j=1}^{e_{ij}} n_{ij}} \qquad (5.5)$$

Assumindo que os n_{ij} são mutuamente independentes, vem:

$$E\left[Y_{ij}\right] = \frac{1}{e_{ij}} \cdot \sum_{i,j=1}^{e_{ij}} E\left[n_{ij}\right] = \frac{1}{e_{ij}} \cdot e_{ij} \cdot m_{ij} = m_{ij} \text{ e } \mathrm{var}\left[Y_{ij}\right] = \frac{1}{e_{ij}^2} \sum_{i,j=1}^{e_{ij}} \mathrm{var}\left[n_{ij}\right] = \frac{1}{e_{ij}^2} \cdot e_{ij} \cdot m_{ij} = \frac{m_{ij}}{e_{ij}}$$

ou seja:

$$\underline{\left| E\left[Y_{ij}\right] = m_{ij} \right|} \qquad (5.6)$$

$$\underline{\left| \mathrm{var}\left[Y_{ij}\right] = \frac{m_{ij}}{e_{ij}} \right|} \qquad (5.7)$$

Assim, a distribuição de Y_{ij} é "semelhante" a uma distribuição de Poisson, com um ajustamento à Variância. Esta abordagem corresponde, no contexto da família exponencial, a utilizar a função densidade de probabilidade na forma definida em *(3.8)*, em que $\phi = 1$ e $\varpi = e_{ij}$.

Alternativamente, poderemos modelar o Número de Sinistros, considerando um Termo Offset igual a $\ln(e_{ij})$ já que, no caso particular de um Modelo de Poisson com função de ligação logarítmica, essa hipótese produz resultados idênticos aos modelados pela via acima (Ver Anexo B).

Na prática, os pressupostos assumidos não se verificam, embora tal não tenha influência na aplicabilidade prática do modelo, como veremos de seguida.

Normalmente, após a ocorrência de um sinistro, a intensidade do risco diminui[63] durante um certo período. Assim, o número de sinistros de cada unidade de exposição não segue uma distribuição de Poisson exacta. Broc*kman e Wrigth* (1992) demonstram que este facto diminui o valor de ϕ, mas que tal não é material, e que, portanto, o modelo em que $\phi = 1$ continua a ser adequado.

Para além disso, pode acontecer que dois ou mais veículos expostos ao Factor de Tarifação i, Nível j estejam envolvidos no mesmo acidente. Esta questão é também abordada em *Brockman e Wrigth* (1992), que demonstram que este facto aumenta o valor de ϕ, mas que tal também não é material.

O terceiro pressuposto foi de que existe homogeneidade em cada Factor/Nível de Tarifação, o que normalmente não acontece na prática, pois a probabilidade de ocorrência de sinistros não é, em geral, a mesma para todas as unidades de exposição desse Factor/Nível.

Por vezes, quando trabalhamos com um modelo de Poisson, a variância da variável resposta é superior ao seu valor médio, isto é $\mathrm{var}[Y] > E[Y]$, ou seja, existe sobre-dispersão. A sobre-dispersão pode ocorrer em várias situações, tal como descrito por *McCullagh e Nelder* (1989). Uma das situações aí descritas é o estudo de propensão a acidentes, em que estudamos uma variável aleatória, W, o número de incidências num determinado indivíduo, com distribuição de Poisson, e em que a própria média dessa distribuição, Z, pode ser vista como uma variável aleatória. Assumindo que Z segue uma distribuição Gama, a realização de W para um determinado indivíduo segue uma distribuição Binomial Negativa.

Assim, no problema em estudo, quando a média do Factor de Tarifação i, Nível j depende da unidade de exposição, verifica-se que o número de sinistros numa determinada unidade de exposição segue uma distribuição Binomial Negativa. Utilizando este resultado, Brokman e Wright (1992) demonstram que a existência de heterogeneidade nos Factores/Níveis de Tarifação pode ser tida em conta considerando que existe sobre-dispersão, ou seja que $\phi > 1$.

Quando existe sobre-dispersão, os erros obtidos também apresentarão sobre-dispersão relativamente à situação teórica. Assim, a Deviance obtida estará aumentada, face à

[63] Por exemplo, por passar a existir mais cuidado por parte do condutor.

Deviance "teórica", por um factor desconhecido, $\phi > 1$. Nesta situação, a comparação da *Deviance* com um quantil $\chi^2_{n-p,\alpha}$, referida em 4.2.3.1., já é menos adequada, sendo que um Teste F é mais robusto. Neste caso, a Estatística F não deve ser comparada directamente com a *Deviance* Residual de um único modelo, sendo apenas aplicável para a comparação de modelos. Ou seja, supondo dois modelos intermédios M_1 e M_2, com M_2 encaixado em M_1 e sendo $D^*(y,\hat{\mu}_j)$ a *Deviance* Residual para o modelo M_j, com j=1,2, então:

$$\frac{D^*(y,\hat{\mu}_2) - D^*(y,\hat{\mu}_1)}{(gl_2 - gl_1)D_1 / gl_2} \sim F_{gl_2-gl_1,gl_1} \quad (5.8)$$

Se o valor da Estatística de Teste for superior a $F_{gl_2-gl_1,gl_1,\alpha}$, então, tal indica que M_2 não é um modelo válido.

5.5. ESCOLHA DA DISTRIBUIÇÃO DA VARIÁVEL RESPOSTA NA MODELAÇÃO DA SEVERIDADE DE SINISTROS

Os primeiros trabalhos relativos à modelação estatística de sinistros incidiram principalmente sobre a modelação da Frequência, enquanto a modelação da Severidade começou a ser abordada mais tarde. Uma das primeiras abordagens a esta questão, foi a modelação da Severidade de sinistros assumindo um Modelo Normal, ou seja, que a variância desta variável é constante, para todos os factores de tarifação[64].

No entanto, a distribuição das Indemnizações é, geralmente, assimétrica, sendo habitualmente ajustável através das distribuições Gama, Pareto, Log-Normal ou Weibul (das quais apenas a distribuição Gama pertence à família exponencial de distribuições). Para além disso, intuitivamente, parece ser mais plausível que, para os factores de tarifação com maior média, a variância seja também superior. De facto, *Coutts (1984),* refuta a assumpção de variância constante, tendo observado que os resíduos da análise acima referida aumentam à medida que as estimativas dos parâmetros aumentam.

McCullagh e Nelder (1989) defendem, para a modelação da Severidade, a escolha da distribuição Gama, com uma parametrização tal que $\mathrm{var}(Y) = \sigma^2 \cdot \mu^2$, que assume que o coeficiente de variação, σ, é constante[65]. Esta hipótese foi seguida posteriormente por *Brockman e Wrigth* (1992).

[64] Baxter, Coutts and Ross (1980)
[65] Note-se que nesta parametrização, σ, não é o desvio padrão.

Assim, com vista a definirmos a componente aleatória do modelo, vejamos que:

$Y \sim Gama(\alpha, \beta)$, com função densidade de probabilidade

$$f(y;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \quad y > 0; \alpha, \beta > 0 \text{ e } \Gamma(\alpha) = \int_0^{\infty} \beta^{\alpha} \cdot y^{\alpha-1} \cdot e^{-\beta y}\, dy \qquad (5.9)$$

Temos que $\mu = E[Y] = \frac{\alpha}{\beta} \Leftrightarrow \beta = \frac{\alpha}{\mu}$

Donde:

$$f(y;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} = \exp\left(\log\left(\beta^{\alpha} y^{\alpha-1} e^{-\beta y}\right) - \log\Gamma(\alpha)\right) =$$

$$= \exp\left(\alpha \log\left(\frac{\alpha}{\mu}\right) + (\alpha - 1)\log(y) - \frac{\alpha}{\mu} y - \log\left(\Gamma(\alpha)\right) \right) =$$

$$= \exp\left(\alpha \log(\alpha) - \alpha \log(\mu) + \alpha \log(y) - \log(y) - \frac{\alpha}{\mu} y - \log\left(\Gamma(\alpha)\right) \right) =$$

$$= \exp\left(\alpha\left(-\frac{y}{\mu} - \log \mu \right) + \alpha \log \alpha y - \log y - \log\Gamma(\alpha) \right) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \right) = f(y \mid \theta, \phi)$$

Ou seja, a distribuição Gama é uma distribuição da família exponencial, em que:

Distribuição **Gama** como membro da Família Exponencial	θ	$b(\theta)$	ϕ	$a(\phi)$	$c(y,\phi)$
	$-\frac{1}{\mu}$	$\log(\mu)$	α	$\frac{1}{\alpha}$	$\alpha \log \alpha y - \log y - \log\Gamma(\alpha)$

5.10 - A Distribuição Gama na Família Exponencial

Temos que:

$$b(\theta) = \log(\mu) = \log\left(-\frac{1}{\theta}\right) = -\log(-\theta) \Rightarrow b'(\theta) = -\frac{1}{\theta} = \mu$$

$$Var\,(Y) = b''(\theta) a(\phi) = -\frac{1}{\theta^2} \cdot \frac{1}{\alpha} = \frac{\mu^2}{\alpha} = \frac{1}{\alpha} \cdot \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2} \text{ e } V(\mu) = \mu^2$$

Na situação em estudo, os dados a modelar terão a seguinte estrutura:

- O número de sinistros ocorridos no Factor de Tarifação i, Nível j, n_{ij}
- A Variável Resposta, Y_{ij}, a Severidade de sinistros, ou seja, o Custo Médio dos n_{ij} Sinistros do Factor de Tarifação i, Nível j
- O custo do k-ésimo sinistro do Factor de Tarifação i, Nível j, Z_{ijk}

Assumindo que Z_{ijk} segue uma distribuição Gama, em que o coeficiente de variação é constante, denotemos:

$E[Z_{ijk}] = c_{ij}$ e $\text{var}[Z_{ijk}] = \sigma^2 \cdot c_{ij}^2$, sendo σ constante,

Donde,

$$\boxed{Y_{ij} = \frac{1}{n_{ij}} \cdot \sum_{i,j=1}^{n_{ij}} Z_{ijk}} \qquad (5.11)$$

Assumindo que os sinistros são independentes, temos, para a Severidade de Sinistros do Factor de Tarifação i, Nível j[66]:

$$E[Y_{ij}] = \frac{1}{n_{ij}} \cdot \sum_{i,j=1}^{n_{i,j}} E[Z_{ijk}] = \frac{1}{n_{ij}} \cdot n_{ij} \cdot c_{ij} = c_{ij}$$

$$\text{var}[Y_{ij}] = \frac{1}{n_{ij}^2} \cdot \sum_{i,j=1}^{n_{i,j}} Var[Z_{ijk}] = \frac{1}{n_{ij}^2} \cdot n_{ij} \cdot \sigma^2 \cdot c_{ij}^2 = \frac{\sigma^2 \cdot c_{ij}^2}{n_{ij}}$$

Ou seja:

$$\underline{E[Y_{ij}] = c_{ij}} \qquad (5.12)$$

$$\underline{Var[Y_{ij}] = \frac{\sigma^2 \cdot c_{ij}^2}{n_{ij}}}, \text{ sendo } \sigma \text{ constante} \qquad (5.13)$$

Então, Y_{ij} possui uma estrutura Gama, com coeficiente de variação constante, e, no contexto da família exponencial, em que se toma $\varpi = n_{ij}$ e $\phi = \sigma^2$. Sendo a função

[66] Pelas propriedades da Média e da Variância

densidade probabilidade escrita na forma definida em (3.8), em que a função $a(\phi)$ toma a forma $\frac{\phi}{\omega}$. A variável ϕ é estimada, como vimos em 4.2.2.2.

5.6. O PRÉMIO DE RISCO

Dado modelarmos a Frequência e a Severidade separadamente, para cada tipo de sinistros, o prémio é obtido multiplicando cada factor relativo à Frequência por cada factor obtido relativo à Severidade para os mesmos factores. Ou seja, o Prémio Puro do Factor de Tarifação i, Nível j, n_{ij} é dado por:

$$\boxed{\text{Prémio Puro}_{ij} = \sum_t \text{Factor Frequência}_{ijt} * \text{Factor Severidade}_{ijt}}, \qquad (5.14)$$

onde t se refere ao Tipo de Sinistros modelados separadamente.

Obtido o Prémio Puro para cada Factor/Nível e Tipo de Sinistro, obtém-se o Prémio de Risco aplicando uma margem de segurança:

$$\boxed{\text{Prémio de Risco}_{ij} = \sum_t \left(\text{Factor Frequência}_{ijt} \cdot (1+\theta F_t)\right) * \left(\text{Factor Severidade}_{ijt} \cdot (1+\theta S_t)\right)} \quad (5.15)$$

Onde θF é a carga de segurança a aplicar à Frequência de Sinistros e θS é a carga de segurança a aplicar à Severidade de Sinistros.

Tal como já referido ao longo deste trabalho, a carga de segurança deverá ter em conta:

- Custos administrativos e de exploração, tais como custos associados à emissão de apólices, à gestão de sinistros e ao pagamento de comissões;
- Ajustamentos à Frequência de Sinistros, nomeadamente no que se refere ao rateio dos sinistros com custos negativos;
- Ajustamentos à Severidade de Sinistros, nomeadamente no que se refere a Sinistros IBNR e IBNER;
- Ajustamentos para Grandes Sinistros.

5.7. CONSTRUÇÃO DE UMA TARIFA STANDARD

Após a obtenção do Prémio de Risco a aplicar em cada Nível de cada Factor de tarifação, *Coutts* (1984) e *Brockman e Wright* (1992) propõem um último passo, útil, para a construção da tarifa. Normalmente, as tarifas existentes apresentam uma estrutura tabelar, em que, para cada Factor de Tarifação, é definido o agravamento ou desagravamento, face à situação base. Assim, para aferir da adequabilidade da estrutura actual, é útil a obtenção de coeficientes na mesma forma. Para além disso, esta última fase da modelação, permite refinar as estimativas obtidas.

O modelo proposto por *Brockman e Wright* (1992), toma como observações o Prémio de Risco obtido através de (5.15), e define-se da seguinte forma:

- As variáveis explicativas são os Factores de Tarifação;
- A variável resposta Y_{ij} possui uma estrutura Gama;
- Os coeficientes finais são ponderados, tendo em conta a exposição de cada Factor/ Nível de Tarifação, ou seja, ϖ define-se como a Exposição do Factor de Tarifação i, Nível j, e_{ij}.

Esta fase final da modelação irá refinar as estimativas obtidas nas fases anteriores da modelação, pois permite considerar interacções entre os dados não consideradas em cada modelação individualmente. Para além disso, caso das modelações iniciais resulte que existem factores utilizados pouco significativos, os mesmos podem ser retirados nesta fase da modelação.

5.8. PARA ALÉM DO PRÉMIO DE RISCO – A MODELAÇÃO DA PROCURA

Uma vez construída uma Tarifa que vá de encontro ao objectivo de manutenção do equilíbrio técnico, coloca-se a questão se a mesma é competitiva. Esse estudo é, em primeira análise, de marketing, mas pode também ser abordado através dos Modelos Lineares Generalizados. *Murphy, Brockman e Lee* (2000) propõem modelos para a modelação da procura, tanto ao nível da captação de apólices novas, como da retenção da carteira. Esta modelação incorpora, para além dos factores de tarifação e do prémio cotado[67], informações como o canal de distribuição, a forma de pagamento, dados sobre a relação comercial do cliente com a seguradora (se tem outras apólices na companhia), dados sócio-demográficos e

[67] O prémio cotado é inserido na modelação quando se pretende modelar a elasticidade dos preços. Caso se pretenda modelar apenas as taxas relativas de captação de apólices por segmento da tarifa, não é necessário incluir este factor como variável explicativa.

informação de mercado[68] e de marketing[69]. No caso da modelação da retenção de carteira, a variação de preços na renovação deve ser também considerada na modelação.

Existem algumas condicionantes de ordem prática à aplicação destes modelos. Em primeiro lugar, e no que se refere particularmente à modelação da captação das apólices novas, as seguradoras podem não ter uma informação completa ou com dimensão suficiente sobre todas as cotações solicitadas[70]. Para além disso, no mercado português, existe uma política de concessão de descontos comercias sobre a Tarifa base, descontos estes que podem ser diferentes, consoante a seguradora e consoante o pedido de cotação em causa. O valor destes descontos, não é, em geral, conhecido, o que pode também condicionar a análise. Tendo em conta estas condicionantes, esta modelação não será abordada na componente prática deste trabalho.

[68] Por exemplo, expectativa de descida ou subida de preços.
[69] Dado que a tarifa se insere num mercado, é necessário incorporar no modelo um "Índice de Preços de Mercado".
[70] No caso das seguradoras apoiadas em canais de distribuição "tradicionais" (genericamente, os mediadores e corretores), na maioria dos casos, a companhia apenas tem conhecimento das cotações efectivamente concretizadas.

6. APLICAÇÃO A UMA CARTEIRA DE RESPONSABILIDADE CIVIL AUTOMÓVEL

Neste capítulo, aplicam-se os Modelos Lineares Generalizados à modelização da estrutura tarifária de uma companhia de seguros Não-Vida. Será modelizada a Frequência de Sinistros, a Severidade de Sinistros e o Prémio de Risco, de acordo com os Modelos definidos no capítulo anterior. Os dados utilizados referem-se à carteira da companhia relativa a veículos Ligeiros, Mistos e Caminhetas de Uso Particular.

Estando definidas as componentes Distribuição da Variável Resposta e Função de Ligação, visto serem aplicados os modelos já definidos na primeira secção do Capítulo 4, faz-se a análise preliminar dos dados, nomeadamente no que se refere às características do Número de Sinistros, do Montante das Indemnizações e das Variáveis Explicativas. Tanto no que se refere ao Número de Sinistros, como ao montante das Indemnizações, poderia ter sido ajustada uma distribuição teórica. No entanto, dado que tal não é necessário para a formulação do modelo, tal como já referido na Subsecção 4.2.1., optou-se por não efectuar este estudo nesta componente prática do trabalho.

Na segunda secção, apresentam-se os resultados da modelação para a estrutura tarifária vigente na companhia, sendo efectuada uma comparação com os valores da tarifa actual.

Finalmente, na terceira secção, apresentam-se os resultados da modelação de uma estrutura tarifária alternativa.

6.1. O MONTANTE DE INDEMNIZAÇÕES

6.1.1. A escolha do período a analisar

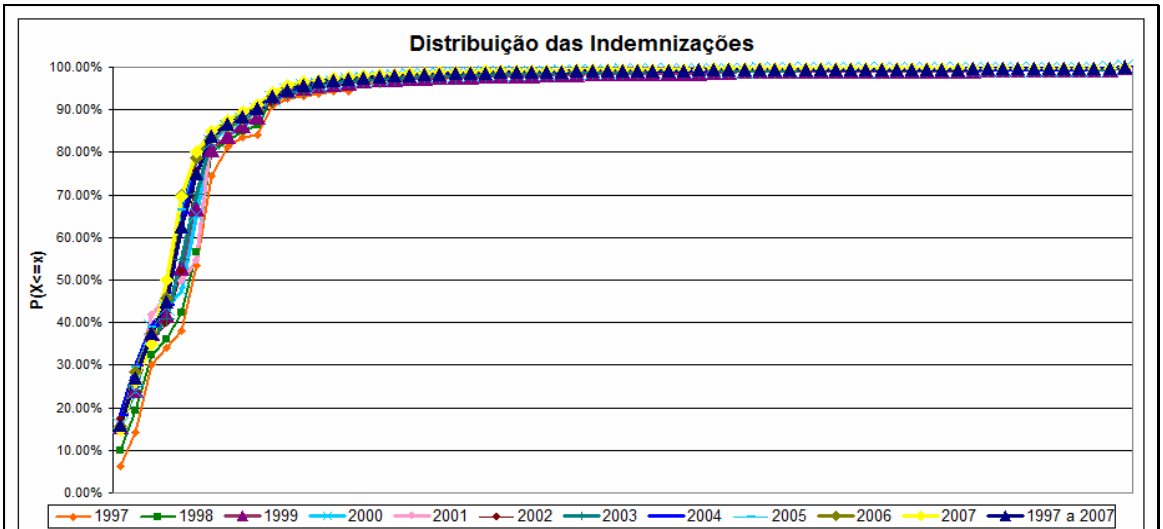
A companhia estudada é uma companhia jovem, e, portanto, com um histórico de sinistralidade ainda relativamente curto. Assim, considerando apenas um período de três anos[71], o volume de dados utilizado seria relativamente curto, pelo que uma análise com base nesse histórico poderia não respeitar os princípios enumerados em 4.2.1.1., nomeadamente no que se refere à fiabilidade e credibilidade dos dados. Por outro lado, interessa-nos utilizar dados homogéneos e que representem a realidade actual.

O ponto de partida desta análise foi o histórico dos sinistros ocorridos entre 1997 e 2007, e abertos até 30-06-2008. De acordo com o já exposto no capítulo 5, consideram-se não só os montantes já efectivamente pagos, como também os montantes ainda em reserva. Os

[71] Ver Subsecção 5.1.1.

sinistros com custos zero não são considerados e os sinistros com custo negativo foram rateados pelos restantes sinistros. Nesta primeira análise, os Grandes Sinistros mantêm-se na base de dados, sem qualquer truncagem de valores.

Como auxílio a esta escolha, foi analisada, como se segue, a função de distribuição empírica, por ano de ocorrência, de forma a verificar se existem anos menos recentes com um comportamento diferente dos anos mais recentes:

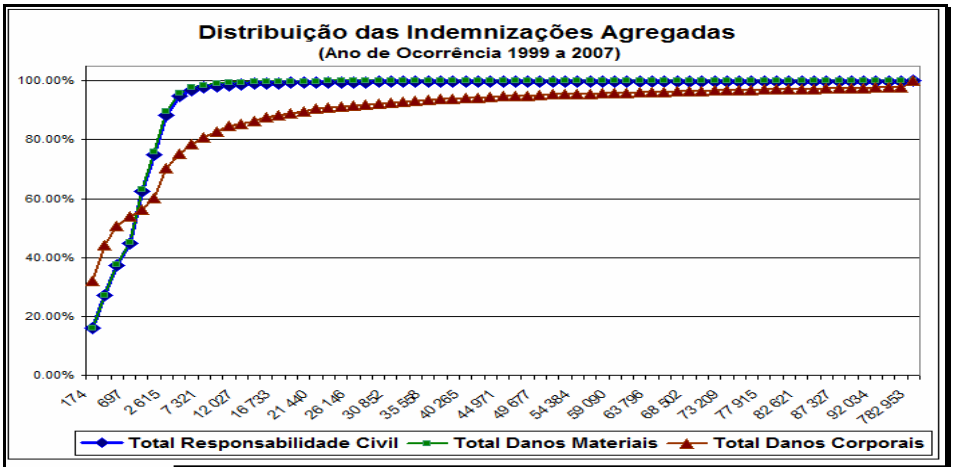


	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	1997 A 2007
Média	2 295	1 956	1 710	1 723	1 640	1 859	1 788	1 636	1 551	1 414	1 340	1 574
Desvio Padrão	8 715	6 594	5 009	9 907	6 874	9 302	9 980	11 010	7 513	7 876	6 850	8 481

Figura 6-1 - Distribuição das Indemnizações por Ano de Ocorrência

Analisando estes indicadores estatísticos, verifica-se que os anos mais recentes têm uma média e desvio padrão mais baixos que os anos menos recentes. Inclusivamente, decrescem, até 2007, ano que apresenta valores mais reduzidos. Ou seja, estes indicadores indiciam que os anos de ocorrência mais recentes têm um "melhor" comportamento, em termos de sinistralidade, que os anos menos recentes. Recorde-se, no entanto, que, à data em que os dados foram retirados, poderíamos ainda não conhecer todos os sinistros de 2007, por poderem ainda vir a ser comunicados sinistros tardios (embora o seu peso deva já ser reduzido). Por outro lado, podem ainda existir desenvolvimentos adversos no provisionamento casuístico de alguns sinistros graves, com valores mais elevados, que poderão alterar alguns indicadores. Pese embora o ano 2007 apresente indicadores diferentes dos restantes, o mesmo foi mantido na análise por ser o ano mais recente e, portanto, o que está mais próximo da realidade que pretendemos tarifar. No entanto, optou-se por excluir da análise os anos 1997 e 1998, dado serem anos menos recentes e, segundo informação da própria companhia estudada, anos menos típicos, dado serem anos com um histórico mais reduzido e irregular.

Quanto à tipologia de sinistros, verifica-se que, considerando o histórico a modelar, a função de distribuição empírica dos sinistros de Danos Materiais é bastante semelhante à que se obtém considerando a totalidade dos sinistros de Responsabilidade Civil, mas a função de distribuição empírica dos sinistros de Danos Corporais segue uma tendência mais diferenciada, com uma média e desvio padrão bastante superiores aos do conjunto dos dois riscos, como seria de esperar:



	Total Responsabilidade Civil	Total Danos Materiais	Total Danos Corporais
Média	1 569	1 135	13 902
Desvio Padrão	8 497	1 866	37 292

Figura 6-2 - Distribuição das Indemnizações por Tipo de Dano

Tal como exposto na Subsecção 5.1.3., interessa-nos identificar os sinistros a considerar como "grandes sinistros". Para tal, foram efectuadas várias simulações, truncando os custos com os sinistros individuais por vários montantes. Considerando os custos assim calculados, calculou-se a média e o desvio padrão. Naturalmente, quanto menor o valor pelo qual os sinistros são truncados, menor o custo médio e o desvio padrão. No entanto, interessa igualmente que o montante total de custos acima desses valores não seja demasiado elevado, pelo que o valor a partir do qual se considera o sinistro como "grande sinistro" teve em conta não só a diminuição da média e do desvio padrão, como também a percentagem de custos truncada. Assim, considera-se "grande sinistro":

- Nos sinistros de Responsabilidade Civil (Danos Materiais e Corporais) – 75.000€
- Nos sinistros de Danos Materiais – 25.000€
- Nos sinistros de Danos Corporais– 200.000€

6.2. O NÚMERO DE SINISTROS

Apresenta-se de seguida uma análise relativa ao número de sinistros, donde se pode verificar que a maioria das apólices expostas não teve sinistros. De facto, em termos globais,

verifica-se que cerca de 80% das apólices expostas[72] não teve sinistros durante o período considerado[73] e que este peso em Danos Corporais (98,88%) é superior ao peso das apólices expostas sem sinistros de Danos Materiais (80,66%).

	Peso número de apólices expostas (*)			Peso do Custo com Sinistros (considerando Sinistros Graves rateados)		
Nº de Sinistros	Responsabilidade Civil	Danos Materiais	Danos Corporais	Responsabilidade Civil	Danos Materiais	Danos Corporais
0	80.45%	80.66%	98.88%	0.00%	0.00%	0.00%
1	15.26%	15.13%	1.10%	62.54%	62.37%	96.69%
2	3.33%	3.28%	0.01%	25.92%	25.91%	2.51%
3	0.74%	0.72%	0.0003%	7.88%	8.16%	0.8015%
4	0.17%	0.16%	0.00%	2.46%	2.47%	-
5	0.03%	0.03%	0.00%	0.57%	0.57%	-
6	0.01%	0.01%	0.00%	0.16%	0.18%	-
Mais de 7	0.00%	0.00%	0.00%	0.47%	0.34%	-
Frequência	3.18%	3.14%	0.14%			

(*) Número de apólice que vigoram entre 1999 e 2007, ainda que tenham sido entretanto anuladas.

Tabela 6.1 - Número de Sinistros

A frequência de sinistros, nos 8 anos em estudo, foi de 3,18% no global, de 3,14% em Danos Materiais e de 0,14% em Danos Corporais.

Na carteira em estudo, a maioria dos sinistros ocorridos, cerca de 77.800, originou custos de Danos Materiais. O Número de Sinistros de Danos Corporais no período estudado é de cerca de 3.500, pelo que, dado que a amostra desta tipologia de sinistros é bastante reduzida, a mesma não será utilizada na modelação. Assim sendo, apenas será modelada a carteira global de Responsabilidade Civil, sem se particularizar a análise por tipo de dano.

6.3. AS VARIÁVEIS EXPLICATIVAS

6.3.1. A Tarifa Actual

A tarifa da companhia estudada tem oito factores de tarifação, cada um com vários níveis. De seguida apresenta-se uma análise preliminar destes factores, com o objectivo de caracterizar a carteira em estudo. Esta análise apresenta algumas estatísticas simples e é uma análise factor a factor que, como já referido, não tem em consideração as interacções entre os vários factores, que podem existir e, desse modo, condicionar as conclusões retiradas. Assim, estas primeiras conclusões, apenas servirão de guia à análise a efectuar nos pontos 6.4. e 6.5., em que será efectuada a modelação conjunta dos vários factores e

[72] Utiliza-se a Exposição Subscrita, ou seja, as apólices expostas ao risco desde o início ao final do período em análise, ainda que tenham sido anuladas antes do fim do período estudado.
[73] Anos de ocorrência entre 1999 e 2007

seus níveis. O nível 1 correspondente à situação "standard", ou seja, o Nível Base desse factor de Tarifação. A codificação do Nível Base com o valor 1, prende-se sobretudo com o facto de o software utilizado na modelação assumir o primeiro nível de cada factor como o Nível Base. Os restantes níveis foram classificados por ordem crescente de gravidade, de acordo com o definido na tarifa em estudo, excepto no que se refere ao Factor Bónus Malus, como se verá de seguida. Em algumas situações, existiam apólices com dados inválidos ou inexistentes, que foram agrupados e considerados como um dos níveis do factor de tarifação.

Para além de se apresentarem os cálculos, por Nível de cada factor, da Frequência, do Custo Médio (que expressa a Severidade de Sinistros, como já referido na Subsecção 2.1.1.) e do peso dos "grandes sinistros" – intervenientes na modelação - apresenta-se ainda o cálculo do Loss Ratio[74]. Para o cálculo deste último indicador, não foi utilizado o conceito de Prémio Bruto Emitido, habitualmente utilizado pelas seguradoras no cálculo do Loss Ratio. Dado que, nas análises seguintes, se pretende estudar a tarifa "integral", sem considerar a aplicação de quaisquer descontos comerciais, considerou-se o Prémio "teórico de tarifa", ou seja, o Prémio calculado de acordo com os coeficientes previstos na tarifa em vigor, para cada Nível de cada factor, sem aplicação de quaisquer descontos comerciais, e multiplicado pelo período de exposição ao risco[75]. Assim, o Loss Ratio que se apresenta é um indicador meramente teórico, dado que, em geral, o mercado segurador concede descontos sobre a sua tarifa base. A utilização deste indicador visa dar uma primeira visão sobre a relação entre o prémio previsto na tarifa e a sinistralidade real.

Dado que não irá ser modelada a carteira por tipologia de sinistros, a análise aqui apresentada é uma análise da carteira global de Responsabilidade Civil.

Temos o primeiro factor, a Escala de Bónus Malus, com 20 níveis, em que o Nível 1 corresponde ao Nível em que não é concedido desconto nem agravamento, daí ter sido classificado como o Nível Base. Os níveis 2 a 10 correspondem aos níveis em que é concedido um desconto, sendo o Nível 10 aquele em que é concedido um maior desconto; e os Níveis 11 a 20 os níveis em que é aplicado um agravamento sobre o prémio base (sendo o Nível 20 o nível mais gravoso). A carteira estudada apresenta uma grande concentração de apólices nos 9 níveis menos gravosos. A escala de Bónus Malus é um factor de tarifação que "actualiza" o valor do prémio, agravando-o quando a apólice regista sinistros numa determinada anuidade, ou seja, a ocorrência de um ou mais sinistros implica a descida na

[74] Rácio entre o Custo com Sinistros e o montante de Prémios.
[75] Ou seja, tendo em conta o período durante o qual as apólices estiveram expostas ao risco.

escala de Bónus Malus para Níveis mais gravosos. Esta escala tem regras de transição[76], em que, consoante o número de sinistros ocorridos na anuidade, a apólice transita na escala, descendo um ou vários níveis, ou seja, agravando o prémio. O funcionamento deste factor de tarifação pode ser constatado na Tabela 6.2., em que se pode verificar que a Frequência, em geral, aumenta, à medida que descemos na escala de Bónus Malus para níveis mais gravosos. No que se refere ao Custo Médio[77] e ao Loss Ratio, estes apresentam montantes mais oscilantes entre os vários níveis, o que é justificado pelo facto de a transição na escala ter em conta o número de sinistros ocorridos na apólice, mas independentemente do custo do sinistro.

Factor de tarifação - Nível BONUS MALUS					
Nível	Peso Nº de apólices expostas	Frequência	Custo Médio	Loss Ratio	Peso Grandes Sinistros
10	61.81%	1.45%	1 171	14%	0.07%
9	8.33%	3.90%	1 652	47%	0.19%
8	9.89%	5.97%	1 625	64%	0.12%
7	5.56%	5.38%	1 719	57%	0.20%
6	5.58%	6.07%	1 726	59%	0.18%
5	2.05%	10.02%	1 840	98%	0.21%
4	2.06%	10.29%	1 876	96%	0.15%
3	1.35%	13.65%	1 853	118%	0.24%
2	1.19%	12.29%	1 783	97%	0.14%
1	1.54%	7.95%	2 399	76%	0.35%
11	0.25%	28.31%	1 613	165%	0.08%
12	0.19%	30.94%	2 157	221%	0.32%
13	0.07%	36.29%	1 462	162%	0.00%
14	0.05%	38.37%	1 656	180%	0.29%
15	0.02%	41.91%	1 149	123%	0.00%
16	0.02%	46.55%	2 155	234%	0.56%
17	0.02%	49.66%	1 403	150%	0.00%
18	0.01%	45.13%	943	85%	0.00%
19	0.00%	46.23%	1 136	93%	0.00%
20	0.01%	58.92%	1 675	157%	0.76%
TOTAL	**100.00%**	**3.18%**	**1 569**	**36%**	**0.15%**

Tabela 6.2 - Análise preliminar dos factores de tarifação: Bónus Malus

O Capital mínimo obrigatoriamente seguro na cobertura de Responsabilidade Civil do seguro automóvel, definido pelo Artigo 12º do Decreto Lei 291/07, de 21 de Agosto, é de 1.800.000€, com sub-limites de 1.200.000€ para sinistros de Danos Corporais e de 600.000€ para sinistros de Danos Materiais. No entanto, as companhias de seguros podem contratar com os seus clientes limites de capital mais elevados, sendo que, no mercado português, o nível

[76] Como já referido, as regras de transição da Escala de Bónus Malus não serão analisadas no âmbito deste trabalho.
[77] Tal como especificado na Subsecção 2.1., a Severidade de Sinistros será expressa pelo Custo Médio.

de capital previsto mais elevado é de 50.000.000€. Em situações em que está envolvida uma terceira entidade, como sejam, por exemplo, as situações de contratos de compra de veículos em regime de *Leasing,* é comum que estas entidades exijam a contratação do seguro de Responsabilidade Civil pelo capital máximo possível de 50.000.000€. Este é o segundo factor estudado, que tem 5 níveis, sendo naturalmente o nível 1, o correspondente ao capital mínimo obrigatório, principal razão pelo qual este factor apresenta uma grande concentração de apólices neste nível - cerca de 96% - como poderá verificar-se da análise da tabela 6.3. O segundo nível com maior concentração de apólices é o que corresponde ao capital máximo contratável, sendo que esta concentração se poderá justificar pelo referido acima relativamente aos contratos em que existe uma terceira entidade envolvida.

No que se refere à Frequência deste factor, a mesma aumenta com a gravidade do nível, excepto no nível 3, o que pode ser explicado pelo reduzido peso das apólices expostas. Quanto ao Custo Médio e ao Loss Ratio, os indicadores mais elevados registam-se no nível menos gravoso e no mais gravoso, ao que não será alheio o facto de estes serem os níveis com um maior peso de apólices.

Factor de tarifação - CAPITAL DE RESPONSABILIDADE CIVIL					
Nível	Peso Nº de apólices expostas	Frequência	Custo Médio	Loss Ratio	Peso Grandes Sinistros
1	95.81%	3.14%	1 561	20%	0.14%
2	0.41%	3.07%	1 390	16%	0.31%
3	0.03%	2.37%	735	6%	0.00%
4	0.26%	3.52%	1 175	13%	0.00%
5	3.50%	4.18%	1 775	22%	0.23%
TOTAL	**100.00%**	**3.18%**	**1 569**	**20%**	**0.15%**

Tabela 6.3 - Análise preliminar dos factores de tarifação: Capital de Responsabilidade Civil

É comum as companhias de seguros tarifarem as suas apólices em função do concelho de circulação habitual do condutor[78]. Em geral, o concelho de circulação é classificado numa Zona, que agrupa os concelhos com níveis de gravidade de risco considerados semelhantes. A companhia estudada considera na sua tarifa três zonas, como constante da Tabela 6.4., em que as mesmas foram classificadas pela ordem de gravidade atribuída pela tarifa. A Frequência aumenta com a gravidade da zona de circulação habitual do condutor, o que não acontece com o Custo Médio, que é menor no nível mais gravoso. Este factor, devido á sua natureza, apresenta maior dispersão de apólices pelos três níveis.

[78] Como vimos na Subsecção 1.2., o Índice de Sinistralidade rodoviária apresenta diferenças consoante a região do país.

Factor de tarifação - ZONA DE CIRCULAÇÃO					
Nível	Peso Nº de apólices expostas	Frequência	Custo Médio	Loss Ratio	Peso Grandes Sinistros
1	28.51%	2.85%	1 564	20%	0.15%
2	29.33%	2.99%	1 790	20%	0.23%
3	42.15%	3.53%	1 439	18%	0.09%
TOTAL	100.00%	3.18%	1 569	19%	0.15%

Tabela 6.4 - Análise preliminar dos factores de tarifação: Zona de Circulação

Um dos factores de tarifação utilizados pela companhia estudada para caracterizar o veículo é a Potência do mesmo, pois assume-se que veículos de potência superior são um risco mais gravoso. A potência do veículo está agrupada em oito escalões, sendo que a maioria das apólices (92,4%) se concentra nos 3 primeiros níveis, que apresentam também indicadores mais elevados, tanto no que se refere à Frequência como à Severidade (neste último caso, o nível 8 apresenta o Custo Médio mais elevado).

Factor de tarifação - POTÊNCIA DO VEÍCULO					
Nível	Peso Nº de apólices expostas	Frequência	Custo Médio	Loss Ratio	Peso Grandes Sinistros
1	52.40%	4.00%	1 779	24%	0.19%
2	19.26%	2.34%	1 289	10%	0.08%
3	20.74%	2.66%	1 291	11%	0.10%
4	6.42%	2.43%	1 258	9%	0.07%
5	1.16%	2.14%	1 203	6%	0.14%
6	0.02%	1.78%	583	2%	0.00%
7	0.00%	2.47%	820	3%	0.00%
8	0.00%	2.22%	3 186	9%	0.00%
TOTAL	100.00%	3.18%	1 569	16%	0.15%

Tabela 6.5 - Análise preliminar dos factores de tarifação: Potência do Veículo

O veículo é também tarifado por um factor Marca, que avalia a perigosidade do veículo, em função de várias características do veículo. Este factor tem três níveis, apresentando uma grande concentração de apólices no nível 1 (cerca de 97%). No que se refere à Frequência, a mesma diminui com a gravidade do nível, sendo que o Custo Médio é também mais elevado no nível menos gravoso, como exposto na Tabela 6.6.

Factor de tarifação - MARCA					
Nível	Peso Nº de apólices expostas	Frequência	Custo Médio	Loss Ratio	Peso Grandes Sinistros
1	96.51%	3.20%	1 568	20%	0.15%
2	2.34%	2.71%	1 618	17%	0.18%
3	1.15%	2.31%	1 571	13%	0.15%
TOTAL	**100.00%**	**3.18%**	**1 569**	**20%**	**0.15%**

Tabela 6.6 - Análise preliminar dos factores de tarifação: Marca

O sexto factor, o Tipo de Veículo, ou seja, genericamente, a classe do veículo, tem três níveis, apresentando também uma grande concentração de apólices no nível 1 (cerca de 84%). O Custo Médio aumenta com a gravidade dos níveis, o que também acontece com a Frequência, excepto no nível 2.

Factor de tarifação - TIPO DE VEÍCULO					
Nível	Peso Nº de apólices expostas	Frequência	Custo Médio	Loss Ratio	Peso Grandes Sinistros
1	84.06%	3.16%	1 526	19%	0.14%
2	9.39%	2.92%	1 736	16%	0.15%
3	6.55%	3.85%	1 840	20%	0.22%
TOTAL	**100.00%**	**3.18%**	**1 569**	**19%**	**0.15%**

Tabela 6.7 - Análise preliminar dos factores de tarifação: Tipo de Veículo

A cobertura Responsabilidade Civil abrange os ocupantes que viajam no veículo, excepto o condutor. Assim, o Número de Lugares do veículo é um factor considerado na tarifação. Este tem cinco níveis, sendo que o nível 5 contém os dados inválidos. A maioria das apólices (cerca de 97%) concentra-se no nível 1, sendo que, em geral, tanto a Frequência como a Severidade aumentam à medida que a gravidade dos níveis aumenta (excepção feita ao nível 3, no caso da Frequência e ao nível 4, no caso da Severidade). Neste caso, os indicadores obtidos, confirmam, em geral, que este factor de tarifação influencia mais o custo médio que a Frequência, o que se esperava, dado que quanto maior o número de sinistrados, maior poderá ser o custo do sinistro (naturalmente dependendo das lesões envolvidas em cada caso).

Factor de tarifação - LUGARES					
Nível	Peso Nº de apólices expostas	Frequência	Custo Médio	Loss Ratio	Peso Grandes Sinistros
1	96.77%	3.17%	1 557	20%	0.14%
2	1.95%	3.62%	1 471	19%	0.18%
3	0.21%	2.56%	3 158	27%	0.80%
4	1.07%	3.63%	2 494	28%	0.44%
5	0.003%	0.00%	-	-	-
TOTAL	**100.00%**	**3.18%**	**1 569**	**20%**	**0.15%**

Tabela 6.8 - Análise preliminar dos factores de tarifação: Número de Lugares

É também comum, no mercado segurador, uma tarifação diferenciada em função da idade e do sexo do condutor habitual. Recorde-se, a este propósito, que, como já foi abordado na Subsecção 1.2., o Índice de Sinistralidade rodoviária apresenta diferenças em função da idade e do sexo do condutor envolvido no acidente. Assim, este factor, apresenta, na tarifa da companhia estudada, cinco níveis, sendo que o nível 5 agrupa os dados inválidos. A maioria das apólices (cerca de 98,5%) concentra-se no nível 1, sendo que, em geral, a Frequência aumenta à medida que a gravidade dos níveis aumenta (excepção feita ao nível 3 e 5). No que se refere à Severidade, a mesma é mais elevada nos níveis 1 e 3.

Factor de tarifação - CLASSE DE IDADE					
Nível	Peso Nº de apólices expostas	Frequência	Custo Médio	Loss Ratio	Peso Grandes Sinistros
1	98.53%	3.16%	1 568	20%	0.15%
2	0.74%	4.33%	1 301	15%	0.00%
3	0.40%	3.94%	1 988	17%	0.25%
4	0.32%	4.61%	1 800	13%	0.26%
5	0.01%	0.00%	-	-	-
TOTAL	**100.00%**	**3.18%**	**1 569**	**20%**	**0.15%**

Tabela 6.9 - Análise preliminar dos factores de tarifação: Classe de Idade

Em resumo, a carteira caracteriza-se por uma elevada concentração de apólices expostas nos níveis referentes aos riscos menos gravosos, excepto no que se refere ao factor de tarifação "Zona", em que há uma maior distribuição das apólices expostas pelos três níveis de tarifação. Seria de esperar que tanto a Frequência como o Custo Médio aumentassem á medida que os níveis de cada factor agravam, o que nem sempre acontece. Apesar de a experiência de sinistralidade de uma seguradora poder evidenciar um comportamento diferente do comportamento teórico assumido pela tarifa da companhia, no caso em estudo, o facto de existirem níveis com poucas apólices poderá influenciar os resultados aqui obtidos. De facto, existindo poucas apólices num determinado nível, poderá acontecer que essas

apólices tenham uma sinistralidade atípica, diferente da que teriam num universo maior e mais diversificado de apólices. Esta concentração de apólices nos Níveis Base poderá influenciar os resultados obtidos na modelação, pelo que os mesmos devem ser analisados tendo esse aspecto em consideração.

Em geral, os grandes sinistros têm um peso reduzido na carteira, sendo que se registaram mais grandes sinistros nos níveis dos factores de tarifação em que se registaram mais sinistros e com maior número de apólices expostas. Assim, parece não existir nenhuma tendência na carteira para a ocorrência de grandes sinistros em determinados níveis dos factores de tarifação.

6.3.2. Factores tarifários alternativos

Numa segunda modelação, irá ainda ser incorporado um outro factor, a Cilindrada do veículo, não considerado na tarifa actual. Este factor será utilizado em substituição do Factor "Potência", pelo que representa uma filosofia de tarifação alternativa. Este factor tem 6 níveis, sendo o nível 6 relativo a dados em que o valor da cilindrada do veículo não consta da base de dados. A maioria das apólices concentra-se nos níveis 2 e 3 (cerca de 87% das apólices). A Frequência agrava com o aumento da gravidade dos níveis, o mesmo acontecendo com o Custo Médio, à excepção do Nível 4, neste último caso, como pode analisar-se na tabela 6.10.

Factor de tarifação - CILINDRADA				
Nível	**Peso Nº de apólices expostas**	**Frequência**	**Custo Médio**	**Peso Grandes Sinistros**
1	6.69%	2.70%	1 505	0.19%
2	71.62%	3.19%	1 527	0.14%
3	15.13%	3.54%	1 763	0.16%
4	0.81%	3.98%	1 566	0.13%
5	0.13%	4.91%	2 239	0.00%
6	5.62%	2.46%	1 565	0.15%
TOTAL	**100.00%**	**3.18%**	**1 569**	**0.15%**

Tabela 6.10 - Análise preliminar dos factores de tarifação: Factor Alternativo - Cilindrada

6.4. MODELAÇÃO DA ESTRUTURA TARIFÁRIA ACTUAL

Partindo dos modelos explicitados no capítulo 5 e com base na amostra seleccionada em 6.1., pretende-se modelar a Frequência e a Severidade de sinistros da carteira em estudo, com o objectivo de estimar os coeficientes a aplicar a cada nível de cada factor da tarifa.

Para esta componente do trabalho, utilizou-se o software estatístico "R"[79]. O formato dos dados a utilizar, bem como os comandos utilizado neste software encontram-se descritos no Anexo D.

Nesta modelação, considera-se o sistema de Bónus Malus como uma variável explicativa, estimando os coeficientes a aplicar a cada classe. De seguida, apresentam-se as estimativas $\hat{\beta}$ obtidas para o Termo de Intercepção e para cada um dos parâmetros, correspondentes a cada um dos níveis. Para que as mesmas sejam comparáveis com os coeficientes da tarifa actual, é necessário calcular $\exp(\hat{\beta})$.

Começamos por modelar o total de sinistros, nas suas componentes Frequência e Severidade. De seguida, calculada a carga de segurança para cada componente, ajusta-se um modelo ao Prémio de Risco, de acordo com o definido nas Subsecções 5.6. e 5.7.

A estimativa obtida para o termo "Intercept", corresponde à estimativa para o Prémio Base, prémio pago por todos os riscos identificados como tendo um factor comum. As estimativas obtidas para cada um dos níveis, correspondem ao factor a multiplicar pelo prémio base. Em termos práticos, como já referido, interessa analisar os coeficientes dados por $\exp(\hat{\beta})$, comparando-os com os da tarifa actual, pelo que se apresenta também esse cálculo e respectiva variação face à tarifa actual.

Como pode observar-se dos resultados expostos na Tabela 6-13, existem níveis que evidenciam um comportamento diferente em cada uma destas duas componentes, Frequência e Severidade, confirmando a ideia subjacente à modelação separada destas. Da análise do erro padrão, verifica-se que algumas estimativas apresentam erros padrão bastante elevados, sendo que as mesmas se referem a níveis do factor de tarifação em que o peso das apólices expostas é reduzido, o que poderá estar a influenciar os resultados. Também se verifica que, estando os níveis de cada factor ordenados por grau "teórico" de gravidade, a evolução dos coeficientes estimados não "confirma" o aumento de gravidade, o que poderá também dever-se ao facto de a maioria das apólices expostas se situar no Nível Base, na maioria dos factores.

[79] R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org

Factor de Tarifação	Nível	Frequência			Severidade			**Prémio de Risco**			Variação face a factor da Tarifa	Peso das apólices expostas
		$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error		
(Intercept)		-2.571	**0.076**	3.4%	7.56992	**1938.985**	6.2%	5.26489	**193.425**	4.1%	-23%	
Bonus Malus	2	0.4654	**1.593**	4.5%	-0.15583	**0.856**	8.2%	0.42483	**1.529**	6.3%	70%	1.19%
	3	0.5512	**1.735**	4.1%	-0.22201	**0.801**	7.5%	0.53271	**1.704**	5.8%	100%	1.35%
	4	0.2825	**1.326**	4.0%	-0.09012	**0.914**	7.3%	0.24071	**1.272**	5.2%	59%	2.06%
	5	0.2543	**1.290**	3.9%	-0.14719	**0.863**	7.2%	0.16825	**1.183**	5.1%	58%	2.05%
	6	-0.2473	**0.781**	3.7%	-0.13873	**0.870**	6.8%	-0.3555	**0.701**	4.5%	0%	5.58%
	7	-0.3721	**0.689**	3.7%	-0.16881	**0.845**	6.8%	-0.5451	**0.580**	4.5%	-11%	5.56%
	8	-0.29	**0.748**	3.5%	-0.1968	**0.821**	6.3%	-0.4639	**0.629**	4.3%	5%	9.89%
	9	-0.7132	**0.490**	3.6%	-0.21235	**0.809**	6.7%	-0.8853	**0.413**	4.3%	-25%	8.33%
	10	-1.727	**0.178**	3.4%	-0.45002	**0.638**	6.2%	-1.9241	**0.146**	4.1%	-71%	61.81%
	11	1.282	**3.604**	5.2%	-0.19827	**0.820**	9.5%	1.27736	**3.587**	10.5%	226%	0.25%
	12	1.369	**3.931**	5.7%	-0.0691	**0.933**	10.5%	1.36602	**3.920**	12.3%	227%	0.19%
	13	1.519	**4.568**	7.4%	-0.30235	**0.739**	13.4%	1.59005	**4.904**	18.1%	277%	0.07%
	14	1.563	**4.773**	8.4%	-0.17713	**0.838**	15.4%	1.57058	**4.809**	21.8%	244%	0.05%
	15	1.637	**5.140**	10.9%	-0.51974	**0.595**	19.9%	1.63912	**5.151**	30.2%	232%	0.02%
	16	1.747	**5.737**	11.4%	-0.0613	**0.941**	20.8%	1.69945	**5.471**	33.3%	222%	0.02%
	17	1.791	**5.995**	10.7%	-0.34575	**0.708**	19.5%	1.67154	**5.320**	32.0%	188%	0.02%
	18	1.673	**5.328**	19.4%	-0.75834	**0.468**	35.3%	1.65736	**5.245**	57.5%	162%	0.01%
	19	1.678	**5.355**	30.5%	-0.54824	**0.578**	55.5%	1.5107	**4.530**	91.1%	101%	0.00%
	20	1.905	**6.719**	13.1%	-0.13164	**0.877**	23.9%	1.75753	**5.798**	43.3%	132%	0.01%
Capital de Responsabilidade Civil	2	0.2449	**1.277**	8.2%	0.01035	**1.010**	14.9%	0.37691	**1.458**	6.3%	34%	0.41%
	3	-0.0196	**0.981**	36.3%	-0.6745	**0.509**	66.2%	0.03907	**1.040**	24.8%	-13%	0.03%
	4	0.3158	**1.371**	9.5%	-0.19896	**0.820**	17.3%	0.3998	**1.492**	7.9%	19%	0.26%
	5	0.355	**1.426**	2.6%	0.04374	**1.045**	4.7%	0.38196	**1.465**	2.3%	12%	3.50%
Zona de Circulação	2	-0.0803	**0.923**	1.4%	0.10287	**1.108**	2.6%	-0.1421	**0.868**	1.1%	-21%	29.33%
	3	0.021	**1.021**	1.3%	-0.06833	**0.934**	2.4%	0.05541	**1.057**	1.0%	-12%	42.15%
Potência do Veículo	2	-0.0027	**0.997**	1.5%	-0.21689	**0.805**	2.8%	-0.1839	**0.832**	1.1%	-29%	19.26%
	3	0.1644	**1.179**	1.4%	-0.17273	**0.841**	2.6%	0.23735	**1.268**	1.1%	1%	20.74%
	4	0.0894	**1.093**	2.4%	-0.19139	**0.826**	4.3%	0.11661	**1.124**	1.7%	-19%	6.42%
	5	0.0247	**1.025**	5.7%	-0.26083	**0.770**	10.5%	0.08055	**1.084**	3.8%	-31%	1.16%
	6	-0.4065	**0.666**	42.1%	-0.99563	**0.369**	77.0%	-0.4963	**0.609**	24.9%	-70%	0.02%
	7	-0.5644	**0.569**	102.7%	-0.61401	**0.541**	187.2%	-0.7283	**0.483**	71.6%	-81%	0.003%
	8	-0.6442	**0.525**	145.3%	0.67572	**1.965**	264.7%	-0.9042	**0.405**	96.0%	-87%	0.002%
Marca	2	-0.023	**0.977**	3.8%	0.08601	**1.090**	6.9%	0.09484	**1.099**	2.8%	5%	2.34%
	3	-0.2102	**0.810**	5.7%	0.13424	**1.144**	10.5%	-0.0559	**0.946**	3.9%	-14%	1.15%
Tipo de Veículo	2	-0.0337	**0.967**	1.9%	0.01214	**1.012**	3.5%	0.21056	**1.234**	1.5%	-1%	9.39%
	3	0.1057	**1.111**	2.0%	0.13305	**1.142**	3.7%	0.13174	**1.141**	1.7%	-19%	6.55%
Lugares	2	0.1084	**1.114**	3.6%	-0.07254	**0.930**	6.6%	0.26792	**1.307**	3.0%	19%	1.95%
	3	-0.1766	**0.838**	13.0%	0.39637	**1.486**	23.8%	0.0412	**1.042**	9.2%	-13%	0.21%
	4	0.0558	**1.057**	5.0%	0.09905	**1.104**	9.1%	-0.0268	**0.974**	4.2%	-25%	1.07%
	5	-14.54	**0.000**	76370.0%	0	**1.000**		-14.449	**0.000**	100.5%	-	0.00%
Classe de Idade	2	-0.0309	**0.970**	5.0%	-0.14377	**0.866**	9.2%	-0.0238	**0.976**	4.6%	-33%	0.74%
	3	0.0005	**1.001**	7.3%	0.21058	**1.234**	13.4%	0.13231	**1.141**	6.4%	-37%	0.40%
	4	0.1667	**1.181**	7.4%	0.14341	**1.154**	13.5%	0.27198	**1.313**	7.1%	-47%	0.32%
	5	-13.84	**0.000**	33770.0%	0	**1.000**		-13.685	**0.000**	45.9%	-	0.01%

Tabela 6.11 - Modelação da estrutura tarifária actual – Estimativas e Erro Padrão

Obtiveram-se os seguintes resultados para a Deviance:

	Deviance Reduzida	Graus de Liberdade	$\chi^2_{n-p;0,05}$
Frequência	2 767.1	3 001	-
Severidade	972.6	2 121	2 229.3
Prémio de Risco	3 502.9	3 001	3 129.6

Tabela 6.12 - Modelação da estrutura tarifária actual - Análise da Deviance

Na modelação da Severidade, a Deviance Reduzida é inferior ao quantil de probabilidade 95% da χ^2, o que, de acordo com o referido em 4.2.3.1., indica que o modelo é adequado. No entanto, no que se refere à modelação do Prémio de Risco, a Deviance Reduzida é ligeiramente superior ao quantil de probabilidade da χ^2, o que pode indicar a não adequação do modelo.

Apesar de a análise da *Deviance* Reduzida ser apenas um mero indicador, tendo em conta que alguns níveis apresentam um erro padrão elevado e que têm poucas apólices expostas, foi efectuada uma recodificação dos níveis, agrupando alguns níveis em alguns dos factores.

Assim, no Factor Capital de Responsabilidade Civil, agrupou-se o Nível 3, aquele que apresenta um erro padrão maior, e o Nível 4, de forma a agrupar os dois níveis com menor peso de apólices expostas – o nível resultante foi designado por 3, na tabela 6-15.

No Factor Potência do Veículo, agruparam-se os Níveis 6, 7 e 8 - passando o novo nível a designar-se por Nível 6 – não só por estes três apresentarem erros padrão elevados, como também por terem pouca expressividade na carteira.

Nos Factores Lugares e Classe de Idade, o Nível 5 (que contém os dados inválidos) foi incluído no Nível 1 de cada factor. Dado que o peso das apólices destes níveis é muito reduzido e, devido ao facto de se tratar de dados inválidos, decidiu-se considerá-los agrupados à situação standard.

Obtiveram-se os resultados constantes da Tabela 6-13.

Factor de Tarifação	Nível	Frequência			Severidade			**Prémio de Risco**			Variação face a factor da Tarifa	Peso das apólices expostas
		$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error		
(Intercept)		-2.573	0.076	3.5%	7.570	1939.218	6.2%	5.263	193.148	3.0%	-23%	
Capital de Responsabilidade Civil	2	0.245	1.278	8.3%	0.010	1.010	14.9%	0.633	1.883	4.6%	73%	0.41%
	3	**0.274**	**1.316**	**9.3%**	**-0.221**	**0.802**	**16.9%**	**0.666**	**1.947**	**5.4%**	**56%**	**0.29%**
	5	0.355	1.426	2.6%	0.044	1.045	4.7%	0.574	1.776	1.7%	36%	3.50%
Potência do Veículo	2	-0.003	0.997	1.6%	-0.217	0.805	2.8%	-0.004	0.996	0.8%	-15%	19.26%
	3	0.164	1.179	1.5%	-0.173	0.841	2.6%	0.156	1.169	0.8%	-7%	20.74%
	4	0.089	1.093	2.4%	-0.192	0.826	4.3%	0.168	1.183	1.2%	-15%	6.42%
	5	0.023	1.025	5.8%	-0.259	0.772	10.5%	0.230	1.259	2.7%	-20%	1.16%
	6	**-0.446**	**0.640**	**38.1%**	**-0.694**	**0.500**	**68.9%**	**-0.215**	**0.807**	**16.7%**	**-60%**	**0.03%**
Lugares	2	0.109	1.115	3.7%	-0.073	0.930	6.7%	0.128	1.136	2.2%	3%	1.95%
	3	-0.176	0.838	13.2%	0.396	1.486	23.8%	-0.075	0.928	6.7%	-26%	0.21%
	4	0.056	1.058	5.1%	0.099	1.104	9.1%	0.118	1.125	3.1%	-20%	1.07%
Classe de Idade	2	-0.031	0.970	5.1%	-0.144	0.866	9.2%	0.012	1.012	3.4%	-30%	0.74%
	3	0.001	1.001	7.4%	0.211	1.235	13.4%	0.024	1.024	4.7%	-43%	0.40%
	4	0.166	1.181	7.5%	0.143	1.154	13.5%	0.194	1.215	5.1%	-51%	0.32%

Tabela 6.13 - Modelação da estrutura tarifária actual dos Níveis agrupados – Estimativas e Erro Padrão

	Deviance Reduzida	Graus de Liberdade	$\chi^2_{n-p;0,05}$
Frequência	2 709.4	2 941	-
Severidade	968.1	2 124	2 222.1
Prémio de Risco	2 881.2	2 941	3 068.3

Tabela 6.14 - Modelação da estrutura tarifária actual com níveis agrupados - Análise da Deviance

Em geral, verifica-se uma diminuição nos Erros Padrão dos vários níveis, embora o erro Padrão do Nível 6 do Factor 4 continue elevado. A *Deviance* Reduzida obtida na modelação do Prémio de Risco diminui, sendo neste caso inferior ao quantil de probabilidade da χ^2, o que é um indicador de que o modelo se ajusta melhor neste caso.

Assim, o resultado da modelação indica que poderão ser feitos ajustes à tarifa em vigor, inclusivamente no Prémio Base, para o Nível 1 das várias variáveis.

No entanto, em alguns factores, as estimativas obtidas resultam num desagravamento do risco em Níveis teoricamente mais agravados. Por exemplo, no caso do factor Potência do veículo, assume-se, em geral, que o risco aumenta com a potência do veículo, mas o resultado do modelo indica uma tendência diferente, seguindo a tendência verificada na análise univariada efectuada na Subsecção 6.3.1. Comercialmente, poderá não ser viável implementar uma tarifa em que o Prémio é desagravado em consequência do aumento da potência do veículo, o que deverá ser tido em consideração. Já no que se refere a outros factores, como por exemplo, o Capital de Responsabilidade Civil e a Classe de Idade, que têm uma filosofia diferente, os resultados obtidos poderão já ser implementáveis, do ponto de vista prático.

Quanto ao agrupamento de Níveis aqui proposto, no caso do Factor Capital de Responsabilidade Civil, dado que o agrupamento dos Níveis 3 e 4 produziu estimativas com erro padrão mais baixo e que o peso na carteira global das apólices neste nível é bastante reduzido (0,03%), a junção destes dois níveis poderá ser uma alteração a implementar. Já no que se refere à junção de níveis efectuadas no Factor Potência do Veículo, embora tal tenha também diminuído o erro padrão das estimativas, comercialmente, poderá ser importante manter este níveis na tarifa, dada a natureza do factor.

A decisão de proceder a ajustamentos à tarifa deverá ter em conta que a análise aqui efectuada incide sobre uma carteira com uma elevada concentração de apólices no Nível 1 dos vários factores, o que pode condicionar as conclusões retiradas.

6.5. MODELAÇÃO DE UMA ESTRUTURA TARIFÁRIA ALTERNATIVA

Como referido na Subsecção 6.3.2., incluiu-se na modelação um factor alternativo, a Cilindrada, não considerado na tarifa actual e que substituirá o Factor Potência do Veículo. Quanto aos Factores Capital de Responsabilidade Civil, Lugares e Classe de Idade, manteve-se a recodificação de níveis efectuada na segunda modelação da secção anterior. Obtiveram-se os resultados constantes da Tabela 6-15.

Factor de Tarifação	Nível	Frequência			Severidade			**Prémio de Risco**			Variação face a factor da Tarifa	Peso das apólices expostas
		$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Error		
(Intercept)		-2.791	**0.061**	3.9%	7.506	**1818.049**	7.9%	4.995	**147.622**	2.9%	-41%	
Bonus Malus	2	0.463	**1.589**	4.4%	-0.160	**0.852**	8.9%	0.534	**1.706**	4.1%	90%	1.19%
	3	0.555	**1.743**	4.0%	-0.233	**0.792**	8.1%	0.659	**1.933**	3.8%	127%	1.35%
	4	0.284	**1.329**	3.9%	-0.098	**0.907**	7.9%	0.329	**1.389**	3.4%	74%	2.06%
	5	0.257	**1.293**	3.8%	-0.169	**0.844**	7.8%	0.231	**1.260**	3.3%	68%	2.05%
	6	-0.239	**0.788**	3.6%	-0.161	**0.852**	7.3%	-0.172	**0.842**	2.9%	20%	5.58%
	7	-0.361	**0.697**	3.6%	-0.215	**0.807**	7.3%	-0.504	**0.604**	2.9%	-7%	5.56%
	8	-0.268	**0.765**	3.4%	-0.251	**0.778**	6.9%	-0.342	**0.710**	2.8%	18%	9.89%
	9	-0.687	**0.503**	3.5%	-0.288	**0.749**	7.2%	-0.707	**0.493**	2.8%	-10%	8.33%
	10	-1.690	**0.185**	3.3%	-0.532	**0.587**	6.7%	-1.966	**0.140**	2.7%	-72%	61.81%
	11	1.276	**3.583**	5.1%	-0.182	**0.834**	10.3%	1.202	**3.327**	6.8%	202%	0.25%
	12	1.358	**3.887**	5.6%	-0.049	**0.952**	11.3%	1.297	**3.658**	8.1%	205%	0.19%
	13	1.496	**4.464**	7.2%	-0.289	**0.749**	14.6%	1.441	**4.223**	11.8%	225%	0.07%
	14	1.549	**4.705**	8.2%	-0.166	**0.847**	16.7%	1.671	**5.316**	14.2%	280%	0.05%
	15	1.601	**4.960**	10.6%	-0.518	**0.596**	21.5%	1.688	**5.408**	19.7%	249%	0.02%
	16	1.733	**5.658**	11.1%	-0.098	**0.907**	22.5%	1.934	**6.916**	21.7%	307%	0.02%
	17	1.767	**5.853**	10.4%	-0.333	**0.717**	21.1%	1.946	**7.002**	20.9%	279%	0.02%
	18	1.649	**5.200**	18.9%	-0.776	**0.460**	38.3%	1.675	**5.337**	37.4%	167%	0.01%
	19	1.632	**5.113**	29.7%	-0.518	**0.596**	60.2%	1.746	**5.730**	59.4%	155%	0.00%
	20	1.868	**6.477**	12.8%	-0.110	**0.896**	25.9%	2.006	**7.433**	28.2%	197%	0.01%
Capital de Responsabilidade Civil	2	0.239	**1.270**	8.0%	-0.003	**0.997**	16.2%	0.609	**1.839**	4.1%	69%	0.41%
	3	0.280	**1.323**	9.0%	-0.222	**0.801**	18.2%	0.686	**1.987**	4.9%	59%	0.29%
	5	0.355	**1.426**	2.5%	0.070	**1.073**	5.1%	0.572	**1.771**	1.5%	35%	3.50%
Zona de Circulação	2	-0.078	**0.925**	1.4%	0.094	**1.099**	2.8%	0.017	**1.017**	0.7%	-8%	29.33%
	3	0.031	**1.031**	1.3%	-0.078	**0.925**	2.6%	0.061	**1.063**	0.7%	-11%	42.15%
Marca	2	-0.039	**0.961**	3.6%	0.068	**1.071**	7.2%	0.054	**1.055**	1.7%	0%	2.34%
	3	-0.259	**0.772**	5.5%	0.104	**1.109**	11.1%	-0.122	**0.885**	2.5%	-20%	1.15%
Tipo de Veículo	2	-0.078	**0.925**	2.1%	-0.029	**0.971**	4.1%	-0.065	**0.937**	1.0%	-25%	9.39%
	3	0.023	**1.023**	2.5%	0.064	**1.067**	5.1%	0.099	**1.104**	1.4%	-21%	6.55%
Lugares	2	0.071	**1.074**	3.6%	-0.089	**0.915**	7.3%	0.050	**1.051**	2.0%	-4%	1.95%
	3	-0.219	**0.803**	12.7%	0.440	**1.552**	25.8%	-0.148	**0.862**	6.0%	-31%	0.21%
	4	0.040	**1.041**	4.9%	0.094	**1.098**	9.9%	0.074	**1.076**	2.8%	-23%	1.07%
Classe de Idade	2	-0.015	**0.985**	4.9%	-0.140	**0.915**	10.0%	-0.078	**0.925**	3.0%	-36%	0.74%
	3	0.004	**1.004**	7.1%	0.252	**1.552**	14.5%	0.050	**1.051**	4.2%	-42%	0.40%
	4	0.165	**1.180**	7.2%	0.182	**1.098**	14.6%	0.100	**1.105**	4.6%	-56%	0.32%
Cilindrada	2	0.255	**1.290**	2.3%	0.035	**1.036**	4.6%	0.280	**1.323**	1.1%	-	71.62%
	3	0.366	**1.442**	2.8%	0.125	**1.133**	5.7%	0.361	**1.435**	1.4%	-	15.13%
	4	0.376	**1.456**	5.8%	0.083	**1.087**	11.9%	0.415	**1.515**	3.4%	-	0.81%
	5	0.530	**1.700**	12.0%	0.422	**1.525**	24.3%	0.507	**1.660**	7.7%	-	0.13%
	6	-0.086	**0.918**	3.3%	0.066	**1.068**	6.7%	-0.081	**0.922**	1.6%	-	5.62%

Tabela 6.15 - Modelação da estrutura tarifária alternativa – Estimativas e Erro Padrão

Obtiveram-se os seguintes resultados para a Deviance:

	Deviance Reduzida	Graus de Liberdade	$\chi^2_{n-p;0,05}$
Frequência	2 376.0	2 974.0	-
Severidade	852.9	2 071.0	2 178.0
Prémio de Risco	3 007.8	2 974.0	3 102.0

Tabela 6.16 - Modelação da estrutura tarifária alternativa - Análise da Deviance

Em geral, obtiveram-se estimativas com erro padrão mais reduzidos que os obtidos para as estimativas obtidas na secção 6.3., como pode observar-se na figura 6.15. Pela figura 6.16, podemos verificar que, tanto no que se refere à modelação da Severidade como à modelação do Prémio de Risco, a *Deviance* Reduzida é inferior ao quantil de probabilidade da χ^2, o que é um indicador de que o modelo se ajusta bem.

Assim, o resultado da modelação indica que o factor Cilindrada pode ser considerado como uma alternativa à estrutura tarifária em vigor, dado que, em geral, as estimativas apresentam um erro padrão mais baixo que o obtido para a modelação da estrutura tarifária em vigor. No entanto, há que ter em consideração que o Nível 6 deste factor representa 5,62% da carteira de apólices expostas, o que poderá condicionar as estimativas obtidas para os restantes níveis deste Factor. Para além disso, é conveniente ter em conta que o factor Potência do Veículo é um factor que permite uma análise do risco mais apurada, sendo por essa razão considerado por algumas companhias do mercado segurador português.

6.6. ENQUADRAMENTO DA TARIFA NA EXPLORAÇÃO TÉCNICA DA COMPANHIA

Da modelação efectuada resulta uma diminuição, face à tarifa em vigor, do Prémio Base, ou seja, do Prémio de Tarifa do Nível 1 dos vários factores de tarifação. Sendo este o Nível que apresenta, em geral, uma maior concentração de apólices, a implementação desta alteração na tarifa traduzir-se-ia numa redução do nível global de prémios, assumindo que se mantém a política de concessão de descontos da companhia. Tal deverá ser tido em conta na decisão de alteração da tarifa, devendo ser ponderado se tal diminuição vai de encontro aos objectivos de produção definidos pela companhia. Por outro lado, como já abordado na Secção 2.2., o nível global de prémios da companhia não deverá comprometer o equilíbrio técnico do ramo, ou seja, para além das despesas já consideradas na carga de segurança, o prémio global deverá ainda ser suficiente para fazer face a outros custos da companhia, tais como os custos de investimentos e o custo de resseguro.

CONCLUSÕES

O mercado segurador português atravessa um período de desafios não só ao nível económico e comercial, transversal à economia em geral, como também se depara com novas exigências ao nível da definição e assumpção de responsabilidades e dos critérios de solvabilidade. Por outro lado, um dos grandes objectivos de uma companhia de Seguros continua a ser a manutenção do equilíbrio técnico. Alcançar este objectivo passa pela definição de uma tarifa técnicamente equilibrada, nunca esquecendo que a mesma deverá ser competitiva. Com vista a este objectivo, interessa retirar da experiência de sinistralidade da companhia toda a informação que permita construir uma tarifa adequada globalmente, que evite situações de anti-selecção, e que indique o prémio adequado a cada cliente.

A abordagem escolhida para este trabalho, os Modelos Lineares Generalizados, é uma abordagem que permite incorporar esses efeitos, através da modelação da Frequência e da Severidade, tendo em conta os factores tarifários e o seu efeito na sinistralidade, bem como as interações entre os vários factores. Estes modelos produzem também informação sobre a bondade do ajustamento. Para além disso, são flexíveis, robustos e de fácil implementação, tendo em conta as ferramentas informáticas disponíveis actualmente.

Na componente prática deste trabalho, obtiveram-se estimativas para os coeficientes a aplicar aos vários níveis da tarifa actualmente em vigor, tendo sido modelada também a hipótese de agregação de níveis com menor exposição das apólices, e cujas estimativas apresentavam um erro padrão maior. Foi também modelada a hipótese de alteração de um dos Factores da tarifa actual, que seria substituído por outro. Em ambas as modelações, obtiveram-se estimativas que indicam que a tarifa actual poderá ser ajustada, ou por via da alteração dos coeficientes dos Factores de Tarifação actual, ou por via de ser considerado um Factor de Tarifação alternativo. No entanto, há que ter em conta que, para a maioria dos Factores de Tarifação, a maioria das apólices se concentra num dos Níveis (o Nível Base), o que pode condicionar as estimativas obtidas para os restantes Níveis. No que se refere ao Factor alternativo considerado, o mesmo apresenta cerca de 5,6% de apólices expostas com dados inválidos, o que poderá igualmente condicionar as estimativas obtidas para os restantes níveis desse factor.

Assim, as estimativas obtidas poderão ser adoptadas numa futura revisão tarifária a efectuar pela companhia em estudo, da qual poderão resultar ajustamentos à sua política de subscrição.

Como proposta de trabalho futuro fica a modelação da procura, em carteiras que possuam informação para tal, nomeadamente, informação suficiente sobre as cotações solicitadas, quer tenham sido efectivadas ou não. Naturalmente que, para esta modelação, seria necessário incorporar o "Índice de Preços de Mercado", que, pelas razões já expostas, não é de fácil obtenção no mercado português.

"Ratemaking is neither pure science nor pure art" (Charles L. Mc Clenahan)

BIBLIOGRAFIA

[1] Andersen, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., Thandi, N. *(2004)*, A Practitioner's Guide to Generalized Linear Models

[2] Baxter, L.A., Coutts, S.M. and Ross, S.A.F. (1980), Applications of linear models in motor insurance, 21st International congress of actuaries, 2, 11.

[3] Brockman, M. J. and Wright, T. S. (1992), Statistical Motor Ratting: Making Effective Use of your data, Journal of the Institute of Actuaries, 119, III, 457 – 543

[4] Coutts, S. M. (1984), Motor Insurance Ratting, an actuarial approach, Journal of the Institute of Actuaries, III, 87 - 148

[5] Dobson, A. J. (2002), An Introduction to Generalized Linear Models, Chapman and Hall, London

[6] Mc Clenahan, C. L. (2001) – Ratemaking, Foundations of Casualty Actuarial Science, 4th Edition, Actuarial Society - Arlington, Virgínia, pp. 75 - 148

[7] McCullagh, P. and Nelder, J.A. (1989), Generalized Linear Models, 2nd Edition, Chapman and Hall, London

[8] Murphy, K.P., Brockman, M. J. and Lee, P.K.W. (2000), Using Generalized Linear Models to Build Dynamic Pricing Systems, Casualty Actuarial Society Forum, Winter, pp. 107 a 140.

[9] Murteira, B. J. F. (1990), Probabilidades e Estatística, Volume II

[10] Pinheiro, A.C.D. (1997), Modelos Lineares Generalizados, uma aplicação ao seguro de responsabilidade civil automóvel

[11] Pitkänen, P. (1975) - Tariff Theory, ASTIN Bulletin, 8:2, pp. 204-228

[12] Portugal, L. (2007), Gestão de Seguros Não Vida

[13] Reis, A. D. E. (2001), Teoria da Credibilidade, ISEG/Cemapre

[14] Turkman, M.A.A. e Silva, G.L. (2000) - Modelos Lineares Generalizados – da Teoria à Prática

[15]" Relatório do Conselho de Administração" - Banco de Portugal (2005, 2006 e 2007)

[16]"Estatísticas do Emprego" - Instituto Nacional de Estatística (2007)

[17] Relatório de Sinistralidade Rodoviária de 2007, do Observatório de Segurança Rodoviária (2007)

[18] Relatório de Mercado, da Associação Portuguesa de Seguradoras (2006 e 2007)

[19] Relatório do Sector Segurador e Fundos de Pensões (2007), do Instituto de Seguros de Portugal

ALGUNS SITES CONSULTADOS

- Instituto de Seguros de Portugal (ISP): www.isp.pt
- Associação Portuguesa de Seguradores (APS): www.apseguradores.pt
- International Actuarial Association (IAA): www.actuaries.org
- Casualty Actuarial Society: www.casact.org
- R Development Core Team : www.R-project.org
- Autoridade Nacional de Segurança Rodoviária: www.ansr.pt

ANEXOS

ANEXO A – MODELOS COMUMMENTE APLICADOS EM DIVERSAS ÁREAS[80]

Natureza dos Dados	Tipo de Situação	Distribuição de Y	Função de Ligação
Contínua	Dados com Variância constante	Normal: $Y_i \sim N(\mu_i, \sigma^2)$	Identidade
Contínua	Dados positivos; Dados em que a variância cresce com a média; Dados em que $Var(Y) = \sigma^2 \mu^2$, e o coeficiente de variação , σ^2 , é constante	Gamma : $Y_i \sim G(\mu, \nu)$	Inversa Logarítmica *(consoante se pretenda modelar os dados na forma aditiva ou multiplicativa)*
Contínua	Dados que representam tempos de vida / Modelos de Sobrevivência	Inversa Gaussiana $Y_i \sim IG(\mu_i, \sigma^2)$	Logarítmica
Discreta	Dados Binários, em que Y tem apenas dois valores possíveis[81]	Binomial: $Y_i \sim B(m, \pi)$	Logit; Log-log complementar:
Discreta	Dados "Politómicos", em que os dados podem tomar um conjunto de valores fixo [82]	Multinomial	Logit; Log-log complementar:
Discreta	Contagens não na forma de proporções e sem valores fixos	Poisson: $Y_i \sim P(\lambda)$	Logarítmica

[80] *Mc Culagah e Nelder* (1989) e *Turkman e Silva* (2000)
[81] Por exemplo, 0 ou 1.
[82] Por exemplo, a classificação de tipos sanguíneos.

ANEXO B – MODELAÇÃO DO NÚMERO DE SINISTROS EM ALTERNATIVA À MODELAÇÃO DA FREQUÊNCIA DE SINISTROS

Pretende-se aqui mostrar que, em resposta ao problema colocado em 5.4., poderemos modelar o Número de Sinistros, considerando um Termo Offset igual a $\ln(e_{ij})$ em alternativa à modelação "directa" da Frequência de Sinistros, considerando a distribuição de Poisson com a função $a(\phi) = \frac{\phi}{\varpi}$, com $\phi = 1$ e $\varpi = e_{ij}$.

Relembremos que, no problema em questão, as observações de Y agrupam-se em i factores de Tarifação e cada observação pode reportar-se a períodos de exposição da apólice diferentes, ou seja, temos:

- A Variável Resposta Y_{ij}, o Número de Sinistros do Factor de Tarifação i, Nível j
- A Exposição do Factor de Tarifação i, Nível j, e_{ij}
- A Frequência do Factor de Tarifação i, Nível j, F_{ij}, com valor esperado fr_{ij}

Mantendo as hipóteses já apresentadas em 5.4. e tomando a função de ligação logarítmica, obtemos:

$\ln(\mu_{ij}) = \ln(fr_{ij} e_{ij}) = \ln(fr_{ij}) + \ln(e_{ij})$, onde $\ln(e_{ij})$ é o termo Offset.

Assim, o modelo fica especificado na seguinte forma:

$$E[Y_{ij}] = \exp\left\{\sum_{i,j} X_{ij}\beta_{ij} + \ln(e_i)\right\} = \exp\left\{\sum_{i,j} X_{ij}\beta_{ij}\right\} * e_i$$

Dado que, $F_{ij} = \frac{Y_{ij}}{e_{ij}}$, no caso particular de um Modelo de Poisson com função de ligação logarítmica, essa hipótese produz resultados idênticos aos modelados pela via apresentada em 5.4.

ANEXO C – MODELOS HABITUALMENTE APLICADOS NA ÁREA SEGURADORA

Y	**Frequência de Sinistros**	**Número de Sinistros**	**Severidade de sinistros**	**Renovação da carteira e captação de apólices novas**
Função de Ligação	$\ln\mu$	$\ln\mu$	$\ln\mu$	$\ln\left(\frac{\mu}{1-\mu}\right)$ [83]
Erro	Poisson	Poisson	Gamma	Binomial
Parâmetro de escala - ϕ	1	1	Estimado (σ^2)	1
Função Variância – $V(\mu)$	μ	μ	μ^2	$\mu(1-\mu)$
Constante ϖ	e_{ij}	1	n_{ij}	1
Termo Offset - ε	0	$\log(e_{ij})$	0	0

Onde:

- e_{ij} , a Exposição do Factor de Tarifação i, Nível j,
- n_{ij} , o número de sinistros ocorridos no Factor de Tarifação i, Nível j
- σ , o coeficiente de variação, que se assume constante
- A Frequência de Sinistros, o Número de Sinistros e a Severidade de Sinistros são modelados assumindo um modelo multiplicativo

[83] A Probabilidade de renovação e/ou de captação de uma apólice nova está entre 0 e 1

ANEXO D – PROGAMAÇÃO EM "R"

O programa "R" *(version 2.6.1 (2007-11-26))*[84] é um software estatístico de distribuição gratuita na web, inicialmente programado por Robert Gentleman and Ross Ihaka (conhecidos por "*R & R*"), do Departamento de Estatística da Universidade de Auckland e, ao longo do tempo, completado por outros nomes[85]. Este software foi escolhido para este trabalho por ter incorporadas funções estatísticas que permitem modelar os dados através dos métodos descritos neste trabalho.

Para utilizar estas funções, será necessário utilizar os "packages" adicionais "stats" e "boot"[86].

Em primeiro lugar, os dados devem ser agrupados num ficheiro de extensão .txt, com a seguinte estrutura:

Factor de Tarifação 1	…	Factor de Tarifação i	Nº de Unidades de Exposição	Nº de Sinistros	Custo com Sinistros
Nível 1	Nível 1	Nível 1			
....	…	....			
Nível j	Nível j	Nível j			

No que abaixo se expõe, a codificação das variáveis utilizadas é a seguinte:

- Fj- Factor de tarifação j, j=1, …, 9[87]. Incluem-se todos os níveis que têm unidades expostas ao risco, sendo o Nível 1, o Nível Base[88].
- ExposA – Exposição Adquirida, ou seja, o número de apólices expostas ao risco, tendo em conta o período de exposição
- NS_RC – Número de Sinistros de Responsabilidade Civil
- Cto_RC – Custo com Sinistros de Responsabilidade Civil[89]

[84] R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org
[85] Ver "Contributers" em http://www.R-project.org
[86] S original by Angelo Canty <cantya@mcmaster.ca>. R port by Brian Ripley <ripley@stats.ox.ac.uk>. (2007). boot: Bootstrap R (S-Plus) Functions (Canty). R package version 1.2-30.
[87] Estes factores devem ser declarados como "Character" no "R"
[88] Este aspecto é importante, pois o "R" assume o primeiro nível dos dados como o Nível Base.
[89] Considerando como custo máximo por sinistro o montante de 75.000€

De seguida, apresentam-se as funções utilizadas no "R" nos vários passos da modelação. Dado que no capítulo 6 já forma referidos os resultados da modelação, os mesmos não são aqui apresentados.

"

```
>#CONSTRUÇÃO DE UMA TARIFA DE RESPONSABILIDADE CIVIL AUTOMÓVEL.

>#MODELAÇÃO DA SINISTRALIDADE ATRAVÉS DE MODELOS LINEARES GENERALIZADOS.

>#FACTORES TARIFÁRIOS DA TARIFA PRATICADA ACTUALMENTE.

>#Importar dados a modelar.

>Dados_TActual<-read.table("Dados Tarifa Actual.txt",header=T)

>attach(Dados_TActual)

>#RESPONSABILIDADE CIVIL - Total: MODELAÇÃO DA FREQUÊNCIA.

>Freq_RC<-glm(NS_RC/ExposA~F1+F2+F3+F4+F5+F6+F7+F8,family=quasipoisson(link="log"),
weights=ExposA)

>summary(Freq_RC)

>#RESPONSABILIDADE CIVIlL - Total: MODELAÇÃO DA SEVERIDADE.

>Sever_RC<-glm(Cto_RC/NS_RC~F1+F2+F3+F4+F5+F6+F7+F8,family=Gamma(link="log"),
      weights=NS_RC)

>summary(Sever_RC)

> qchisq(0.95, 2121, ncp=0, log = FALSE)

>#RESPONSABILIDADE CIVIL - Total: MODELAÇÃO DO PRÉMIO PURO.

>PrPuro_RC<-
glm(fitted.values(Freq_RC)*(1.28)*fitted.values(Sever_RC)*(1.26)~F1+F2+F3+F4+F5+F6+F
7+F8,family=Gamma(link="log"),weights=ExposA)

>summary(PrPuro_RC)

> qchisq(0.95, 3001, ncp=0, log = FALSE)

>#FACTORES TARIFÁRIOS DA TARIFA PRATICADA ACTUALMENTE - NÍVEIS COM MENOR EXPOSIÇÃO
AGRUPADOS.

>#Importar dados a modelar.

>Dados_TActual2<-read.table("Dados Tarifa Actual-Níveis Agrup.txt",header=T)

>attach(Dados_TActual2)

>#RESPONSABILIDADE CIVIL - Total: MODELAÇÃO DA FREQUÊNCIA.

>Freq_RC<-glm(NS_RC/ExposA~F1+F2+F3+F4+F5+F6+F7+F8,family=quasipoisson(link="log"),
weights=ExposA)

>summary(Freq_RC)
```

```
>#RESPONSABILIDADE CIVIlL - Total: MODELAÇÃO DA SEVERIDADE.

>Sever_RC<-glm(Cto_RC/NS_RC~F1+F2+F3+F4+F5+F6+F7+F8,family=Gamma(link="log"),
      weights=NS_RC)

>summary(Sever_RC)
> qchisq(0.95, 2124, ncp=0, log = FALSE)

#RESPONSABILIDADE CIVIL - Total: MODELAÇÃO DO PRÉMIO PURO.

>PrPuro_RC<-
glm(fitted.values(Freq_RC)*(1.28)*fitted.values(Sever_RC)*(1.26)~F1+F2+F3+F4+F5+F6+F
7+F8,family=Gamma(link="log"),weights=ExposA)

>summary(PrPuro_RC)

> qchisq(0.95, 2941, ncp=0, log = FALSE)

>#FACTORES TARIFÁRIOS NUMA TARIFA ALTERNATIVA.

>#Importar dados a modelar.

>Dados_TAlternativa<-read.table("Dados Tarifa Alternativa.txt",header=T)

>attach(Dados_TAlternativa)

>#RESPONSABILIDADE CIVIL - Total: MODELAÇÃO DA FREQUÊNCIA.

>Freq_RCAlt<-
glm(NS_RC/ExposA~F1+F2+F3+F5+F6+F7+F8+F9,family=quasipoisson(link="log"),
weights=ExposA)

>summary(Freq_RCAlt)

>#RESPONSABILIDADE CIVIlL - Total: MODELAÇÃO DA SEVERIDADE.

>Sever_RCAlt<-glm(Cto_RC/NS_RC~F1+F2+F3+F5+F6+F7+F8+F9,family=Gamma(link="log"),
      weights=NS_RC)

>summary(Sever_RCAlt)

> qchisq(0.95, 2076, ncp=0, log = FALSE)

>#RESPONSABILIDADE CIVIL - Total: MODELAÇÃO DO PRÉMIO PURO.

>PrPuro_RCAlt<-
glm(fitted.values(Freq_RCAlt)*(1.28)*fitted.values(Sever_RCAlt)*(1.26)~F1+F2+F3+F5+F
6+F7+F8+F9,family=Gamma(link="log"),weights=ExposA)

>summary(PrPuro_RCAlt)

> qchisq(0.95, 3023, ncp=0, log = FALSE)
```